2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

EFFICIENT MODELING BY SELECTING LEARNING SAMPLES IN HUMAN POSE ESTIMATION

Norimichi Ukita, Yoichi Matsuyama, and Norihiro Hagita

Nara Institute of Science and Technology

ABSTRACT

While the more learning data the better the recognition, increase in the data causes an expensive computational cost in learning. This paper proposes how to decrease the computational cost by appropriately selecting the learning data. In particular, we put our focus on learning for human pose estimation in still images. Three kinds of methods are proposed for learning data selection in this paper. The first one divides all data into several clusters in a feature space for avoiding duplication of similar data. The second one selects the data based on their distance from a discriminant plane for efficiently updating it. Third one merges those two methods as well as pruning in optimized pose search. Experimental results show that the proposed method can decrease the learning time by 79 % with less decrease in pose estimation accuracy.

Index Terms— Efficient learning, Selecting samples, Human pose estimation, Deformable part model, Latent SVM

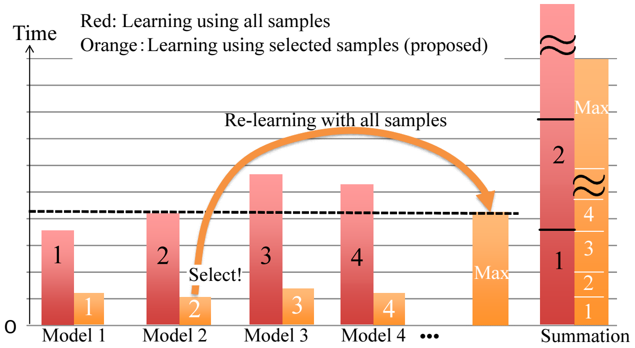

Fig. 1. Effects of reducing a computational cost by the proposed method. Red and orange bars show the computational costs of a general approach (i.e. learning all samples) and the proposed method, respectively.

1. INTRODUCTION

Stochastic models for recognition are improved in general by learning more training samples. A large amount of training samples, however, leads to a huge computational cost. For example, ImageNet [1] has 1.2 million images of 1000 object classes for general object recognition. Model learning with this dataset requires at least 250 days by traditional image features and classifiers [2]. If such long-term model learning is required only once, it might be acceptable. This assumption is not reasonable in research and development purposes because various possible sets of models are tested for obtaining good recognition results. To this end, a computational cost must be suppressed as low as possible.

The proposed method reduces a computational cost by efficiently decreasing training samples. Since random selection of samples degrades recognition accuracy, careful selection of samples is important. By selecting samples that are effective for discrimination between recognition classes, even a small number of samples would be enough for proper recognition.

In sample selection, a trade-off between a computational cost and recognition accuracy is crucial. In trials for model selection, absolute accuracy is not necessarily required but relative accuracy of different models should be verified. Assume that, both in learning all samples and selected samples, the same model gets the maximum accuracy. Then the best model having the max accuracy in trials can be re-trained by all samples. While that model is trained twice in this scheme, the total learning time can be reduced as learning potential models (i.e. Model1, Model2, ..., in Fig. 1) with selected samples (orange bars in the figure) is much faster than that with all samples (red bars in the figure). As a result, the total cost with the selected samples is lower than that with all samples, as shown in “Summation” in the figure.

In this work, it is assumed that the best models in learning selected samples and all sample are identical if recognition accuracy of the model in learning the selected samples is higher. Based on this assumption, we achieve efficient learning and high accuracy simultaneously by appropriate sample selection.

2. RELATED WORK

In this work, the proposed method is applied to human pose estimation in still images. This is because 1) human pose estimation has a variety of applications for observing human

978-1-4799-3694-6/14/$31.00 ©2014 IEEE

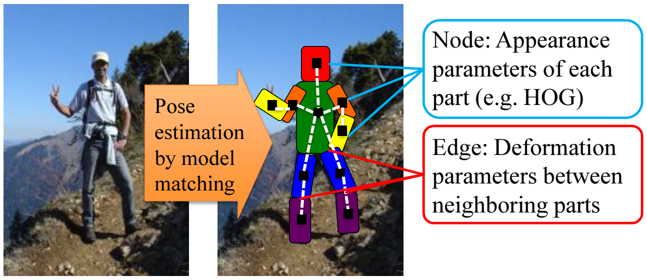
Pose estimation by model matching
Node: Appearance parameters of each part (e.g. HOG)
Edge: Deformation parameters between neighboring parts

Fig. 2. Pose estimation by PSM. Each node and link depicts a part and connection between parts, respectively.

activities and 2) human pose estimation is a complex problem with a huge number of parameters. Since modeling a number of parameters a requires large computational cost, efficient modeling such as the proposed method is useful.

For human pose estimation, appearance features, geometric configuration models such as deformable part models, and discriminative models take key roles.

For representing the appearance of body parts, HOG [3] is powerful and its extensions have been also proposed. For example, the histogram bins of HOG are weighted for body parts detection in [4]. Robustness to occlusion is improved by integrating HOG and LBP [5] in [6]. Appearance between parts can be also modeled [7].

Among all of deformable part models, pictorial structure models (PSM) [8] are most efficient in pose estimation. Other structures have been also proposed for improving ability of expressing the relationship among parts (e.g. [9]). Efficient modeling can be also achieved by clustering parts' appearances [4, 10] and configuration [11, 12].

Discriminatively-trained deformable part models [13] allow us to improve discriminativity for finding correct human poses by learning not only positive (i.e. human regions) but also negative samples.

As summarized above, various parameters (e.g. weights of histogram bins, the number of clusters) are used in models for pose estimation. These parameters should be determined properly for better models.

Efficient learning of human pose models has been proposed in [14, 15]. These previous methods, however, achieve efficient learning only for modeling appearance parameters (i.e. no efficient learning for geometric relationships between parts). In addition, no explicit criterion is used for efficient sample selection. Compared with these previous methods, the proposed method selects samples based on all parameters with several criteria for efficient learning.

3. PICTORIAL STRUCTURE MODELS FOR HUMAN POSE ESTIMATION

A tree-structured PSM is visualized in Fig. 2. The tree model $G = (V, E)$ consists of n nodes $V = \{v_1, ..., v_n\}$, corresponding to body parts, and links $e_{i,j} \in E$, where $e_{i,j}$ connects two parts v_i and v_j. Each node i has its pose parameters, l_i, that localize the respective part. The pose parameters are optimized by minimizing the score function below:

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right), \qquad (1)$$

where $m_i(l_i)$ and $d_{ij}(l_i, l_j)$ denote dissimilarity scores of the appearance feature of i-th part and the deformation score between parts i and j, respectively.

4. MODEL LEARNING BY LATENT SVM AND ITS PROBLEM IN COMPUTATIONAL COST

Discriminative learning for human pose estimation needs "**positive images** where the parts of a human body are annotated" and "**negative images** where no people is observed". In what follows, "the region of a human body observed in a positive image" and "any region observed in a negative image" are called a **positive sample** and a **negative sample**. With positive samples, the appearance features of parts and the geometric relationship between them are modeled as parameters in Eq. 1. Learning negative samples allows us to optimize the parameters so that they discriminate between body parts and other objects that are similar to those parts.

For learning negative samples, the region of a human body in a negative image is detected by a PSM. The detected region must be a false-positive. The detected false-positive is used as a negative sample for re-learning the PSM. In the proposed method, this learning is performed by the latent SVM [13]:

Step1 PSM is trained by the latent SVM with all positive images and a few negative images.

Step2 PSM finds false-positives in a new negative image.

Step3 This false-positive is used as a negative sample by the Latent SVM for re-learning PSM.

Step4 Steps 2 and 3 are repeated until no false-positive is detected.

Step5 If no false-positive is detected, go back with a new negative image to Step 2

The steps 2, 3, 4, and 5 are repeated for all negative images. Negative samples detected in a negative image must be trained one by one with the latent SVM for semi-convex optimization. These iterative learning steps lead to a huge computational cost in learning by the latent SVM.

Table 1. Change in accuracy among different models.

Accuracy (%)	Model A	Model B	Model C
All negative images	**62.67**	62.22	61.94
Randomly-selected images	53.38	**53.70**	53.60

5. EFFECTS OF DIFFERENT SELECTIONS OF SAMPLES

This section proves the necessity of efficient sample selection for human pose estimation using PSM (mentioned in Sec. 3) and discriminative learning (mentioned in Sec. 4).

Given different three models A, B, and C, all of which were trained with the same number of samples, the models were used for human pose estimation. Model A was a mixture-part model [4]. Model B was same with Model A except that the histogram bins of HOG was reduced to 27; Model A had 32 bins. The number of mixture-parts in Model C was reduced to 3; Model A had 5 mixture-parts.

Positive and negative images were given by Leeds Sports Pose (LSP) dataset and INRIA Person dataset, respectively. 2000 images in LSP were divided to 1000 training positive images and 1000 test images. All of 1218 background images in INRIA Person were used as negative sample images. For all models, while the same set of positive images was used, two different sets of negative images were tested. One set consists of all negative images ("All negative images" in Table 1). The other one had 3 % of all negative images, which were selected randomly ("Randomly-selected images" in Table 1). Results shown in all tables in this paper (Tables 1 to 6) were the means of ten trials.

Table 1 shows that the best models obtained from all negative samples and randomly-selected samples were different. These results demonstrated the necessity of efficient sample selection.

6. NEGATIVE SAMPLE SELECTION AND EXPERIMENTS

6.1. Overview

Iterative learning of negative samples is dominant in learning by the latent SVM in terms of a computational cost. This iteration is reduced by efficiently selecting negative samples.

This paper proposes the three selection methods below:

Clustering: Efficiently distributed negative samples are selected by clustering (Sec. 6.3).

Distance from a decision boundary: A decision boundary is updated efficiently by selecting samples that are far from the decision boundary (Sec. 6.4).

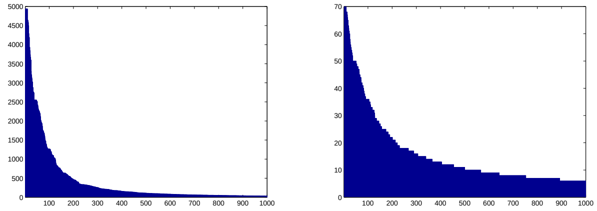

Fig. 3. (Left) False-positives extracted in each negative image. (Right) The number of selected negative samples, which is defined by Eq. (2). Horizontal and Vertical axes show respectively the IDs of negative images and the number of false-positives/selected-samples.

Clustering after distance-based selection: In addition to integration of clustering and distance-based selection, 1) the dimension of a feature vector is decimated for fast clustering and 2) pruning makes pose optimization efficient. (Sec. 6.5).

As with experiments in Sec. 5, LSP and INRIA Person datasets were used in experiments in this section.

6.2. The Number of Negative Samples

The proposed methods determine the number of negative samples depending on the number of false-positives. The relationship between two values is determined in advance based on the propensity of the number of false-positives in negative images given one by one. Figure 3 (Left) shows the numbers of false-positives detected in negative images. The negative images are used for learning in order of their IDs, which appear in the horizontal axis of the graphs in Fig. 3. It can be seen that many false-positives were detected at the beginning of the learning step because of a model is premature.

In accordance with the number of negative samples, the number of negative samples is expressed as follows (shown in Fig. 3 (Right)):

$$N_{NS} = \frac{N_{FP}}{\sqrt{N_{FP}}+1}, \qquad (2)$$

where N_{FP} denotes the number of false-positives. In all experiments shown in the following sections, N_{NS} false-positives are selected for negative samples in each image.

6.3. Proposed Method 1: Selection using Clustering

The proposed method 1 assumes that widely-distributed samples can efficiently improve a human pose model. Based on this assumption, K-means clustering is applied to all false-positives detected in each negative image. A false-positive that is closest to the centroid of false-positives in each cluster

Table 2. (Proposed method 1) Results of pose estimation with negative samples selected by clustering. Each value in brackets indicates the percentage of the cost of learning all samples.

	Accuracy %	Learning Sec	Search Sec	Clustering Sec	Total Sec
Clustering	57.76	653 (11)	1676 (71)	3791	6120 (74)
Random	55.32	714 (12)	1674 (71)	-	2388 (29)
All samples	62.71	5911	2362	-	8273

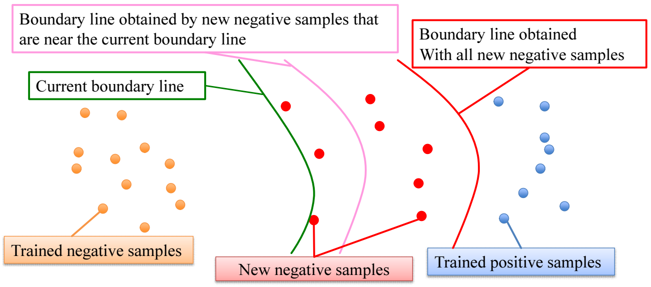
Boundary line obtained by new negative samples that are near the current boundary line
Boundary line obtained With all new negative samples
Current boundary line
Trained negative samples
New negative samples
Trained positive samples

Fig. 4. Selecting negative samples that are far from a decision boundary for efficient learning. By learning only samples that are far from a decision boundary, a red boundary is learned without temporally learning a pink boundary.

is regarded as its representative sample. The number of the clusters is N_{NS} in Eq. 2.

The results of the proposed method 1 are shown in Table 2. The results of all false-positive learning and randomly-selected sample learning are also shown for comparison. In the random selection method, N_{NS} false-positives were selected in each negative image. “Search” in the table shows the computational time for finding all false-positives. The time for learning negative samples is shown in “Learning”. The total computational time is shown in “Total”.

The proposed method improved accuracy 2 % compared with random selection, while the computational cost was reduced to 11 % of that of learning all samples. However, the total time of the proposed method was much larger than that of random selection due to a huge cost for clustering. This clustering was computationally expensive because 1) a large number of false-positives are clustered and 2) the dimension of a false-positive feature vector is huge, 13489 dimensions in the implementation of [4]: 512D HOG at each part, 26 parts, 4 geometric parameters at each link, 25 links, and 77 other parameters $((512 \times 26) + (4 \times (26 - 1)) + 77 = 13489)$.

6.4. Proposed Method 2: Selection based on a Distance from a Decision Boundary

The proposed method 2 selects negative samples each of whose distance from a decision boundary is larger. It is expected that this criterion allows us to significantly update the decision boundary for efficient learning, especially when a human pose model is not matured. In a typical toy example illustrated in Fig. 4, false-positives depicted by red points are detected. If false-positives near a current decision boundary, depicted by a green curve, are selected for negative samples, the decision boundary is updated to a pink curve. When other remaining false-positives are trained later by the latent SVM, a red curve is obtained eventually as the decision boundary. This example gives us intuition about the advantage of learning samples that are far from a decision boundary.

The basic learning steps of the proposed method 2 are same with those described in Sec. 4, except that the distance values between all false-positives and a decision boundary are sorted in descending order for selecting N_{NS} negative samples in Step 3. The following two features are used for computing a distance between a sample and the decision boundary:

Feature of a full body A 13489 dimensional feature vector of a full body is used for computing the distance. This distance is automatically obtained in human pose detection using the PSM trained by the latent SVM.

Appearance feature of a part having the max variance: The variance of each component of HOG is computed for obtaining the summation of the variance values of all components in each part. Given p-th part having the max summation, the 512D HOG feature of p-th part is used for distance computation. This criterion is designed for selecting a part with a variety of appearance for efficient learning. A distance in 512 dimensions is expressed by dissimilarity of p-th part. This dissimilarity score is computed in the pose estimation process ($m_i(l_i)$ in Eq. (1)).

The above two kinds of distance are computed in the pose estimation process, which is performed before negative sample selection. This results in fast sample selection with no additional processes.

Table 3. (Proposed method 2) Results of negative sample selection based a distance from a decision boundary (using LSP + INRIA).

	Accuracy %	Learning Sec	Search Sec	Total Sec
Full body	58.83	772	1683	2465
Max var	57.84	1051	1664	2715
Random	55.81	754	1674	2428
All samples	62.71	5911	2362	8273

Table 4. (Proposed method 2) Results of negative sample selection based a distance from a decision boundary (using PARSE + INRIA).

	Accuracy %	Learning Sec	Search Sec	Total Sec
Full body	70.94	588	2563	3051
Max var	69.89	859	2479	3338
Random	68.24	590	2612	3102
All samples	76.21	3428	4233	7661

We conducted experiments with the two distance metrics. Table 3 shows that 1) the two distance metrics outperformed all-sample learning in terms of accuracy and 2) their computational costs were comparable with that of random selection.

Since the results of the two metrics differed by only 1 % in terms of accuracy, more experiments were conducted with another dataset for further verification. The Image Parse dataset was used as positive and test images. The results are shown in Table 4. The results of the Image Parse dataset also prove the properties of the proposed method 2, which is almost same with those shown in Table 3.

6.5. Proposed Method 3: Selection based on Distribution of Samples

The proposed method 1 requires a huge cost due to clustering huge-dimensional features. The proposed method 2 might select false-positives that are closer to each other, which cause inefficient learning. For complementarily integrating those two methods, the proposed method 3 narrows down false-positives based on a distance from a decision boundary and then clusters them.

The number of negative samples (i.e. the number of clusters) is N_{NS} in Eq. (2). The N_{NS} negative samples are selected by clustering N_{SC} negative samples that are selected from all negative samples by the proposed method 2. In the proposed method, this number N_{SC} is defined as follows:

$$N_{SC} = \frac{N_{FP}}{\sqrt[4]{N_{FP}} + 1} \tag{3}$$

Clustering N_{SC} false-positives is made more efficient by significantly simplifying HOGs in a feature vector. In the base model [4], each part consists of 4×4 cells, each of which has 32 orientation bins: $4 \times 4 \times 32 = 512$ dimensions. The simplified HOG consists of one cell having 16 orientation bins. It is concatenated with 4D vectors, each of which represents the geometric relationship between two parts. In our implementation with 26 parts, $(16 \times 26) + (4 \times 25) = 516$ dimensional feature vectors are clustered. Note that the simplified features are used only for this clustering, while human pose estimation is performed with full-dimensional features.

In addition, a computational cost for searching false-positives is reduced by pruning in pose estimation by PSM. Since PSM employs the dynamic programming for computing the dissimilarity score of each human pose, adding dissimilarity scores in a tree model can be quit when the summation of the scores is larger. A threshold for this pruning is determined with a false-positive having the minimum score, L_{min} defined by Eq. 1, detected in the last iteration. Since L_{min} might decrease by iterative learning of negative samples, the threshold is determined to be $0.9L_{min}$.

The experimental results of the proposed method 3 are shown in Table 5. The distance metric using the full-body feature was used because it was more accurate as shown in Tables 3 and 4. For comparison, the results of the proposed methods 1 and 2 are also included in the table. Recognition accuracy of the proposed method 3 was improved compared with the proposed methods 1 and 2. The computational cost could be also reduced. In total, the computational cost was reduced to 28 % of learning all samples: 8273 sec → 2327.

Table 6. Accuracy in different models.

	Model A %	Model B %	Model C %
All samples	62.71	62.22	61.94
Proposed 2 (Full body)	56.82	56.33	56.19
Proposed 2 (Max var)	55.72	55.19	54.63
Proposed 3 (Full body)	59.82	59.11	58.88
Random	53.38	53.70	53.60

6.6. Different Models learnt by the Proposed Method

We conducted experiments for proving that the proposed methods can get the best model that is obtained also by learning all samples. Only the proposed methods 2 and 3 were tested because their computational costs are much smaller than that of the proposed method 1. All experimental conditions in the experiments were equal to those in Sec. 5.

As shown in Table 6, the best models obtained by learning all samples and the proposed methods were same; model A got the best accuracy in all learning methods excepting random selection. These results empirically demonstrated the effectiveness of the proposed methods.

Table 5. (Proposed method 3) Results of negative sample selection based on a distance from a decision boundary and clustering.

	Accuracy %	Learning Sec	Search Sec	Clustering Sec	Total Sec
Proposed 3 (Full body)	59.82	623 (11)	1068 (58)	46	1737 (21)
Proposed 1 (Only clustering)	57.76	653 (11)	1676 (71)	3791	6120 (74)
Proposed 2 (Full body)	58.83	772 (13)	1683 (71)	-	2465 (30)
Proposed 2 (Max var)	57.84	1051 (18)	1664 (70)	-	2715 (33)
Random	55.32	714 (12)	1674 (71)	-	2388 (29)
All samples	62.71	5911	2362	-	8273

7. CONCLUDING REMARKS

This paper proposed efficient model learning by appropriate sample selection. In a human pose estimation problem, computationally-dominant negative-sample learning is achieved with the reduced number of samples selected by the proposed method. For efficient sample selection, the following three methods are proposed: 1) selection using clustering, 2) selection based on a distance from a decision boundary, and 3) fusion of 1st and 2nd methods with dimensionality reduction of features and pruning in pose optimization. In learning human pose models with 1000 sample images, the computational cost decreased by 79 %.

Future work includes 1) how to guarantee the trade-off between a computational cost and recognition accuracy (i.e. how to determine the number of selected samples in order to guarantee that relative recognition accuracy among models is not changed between learning with all and selected samples), 2) sample selection by more aggressively using the properties of human-pose features, and 3) using the proposed methods for other recognition problems.

8. REFERENCES

[1] Jia Deng, Alexander C Berg, Kai Li, and Fei-Fei Li, "What does classifying more than 10,000 image categories tell us?," in *ECCV*, 2010.

[2] Yuanqing Lin, Fegjun Lv, Shenghuo Zhu, Ming Yang, Timothee Cour, Kai Yu, and Liangliang Cao, "Large-scale image classification:fast feature extraction and svm training," in *CVPR*, 2011.

[3] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[4] Yi Yang and Deva Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011.

[5] Li Wang and Dong Chen He, "Texture classification using texture spectrum," *Pattern Recognition*, vol. 23, no. 8, pp. 905–910, 1990.

[6] Xiaoyu Wang, Tony X. Han, and Shuicheng Yan, "An hog-lbp human detector with partial occlusion handling," in *ICCV*, 2009.

[7] Norimichi Ukita, "Articulated pose estimation with parts connectivity using discriminative local oriented contours," in *CVPR*, 2012.

[8] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[9] Jiayong Zhang, Jiebo Luo, Robert T. Collins, and Yanxi Liu, "Body localization in still images using hierarchical models and hybrid search," in *CVPR*, 2006.

[10] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *ICCV*, 2013.

[11] Sam Johnson and Mark Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *BMVC*, 2010.

[12] Sam Johnson and Mark Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*, 2011.

[13] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan, "Discriminative latent variable models for object detection," in *ICML*, 2010.

[14] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009.

[15] Vivek Kumar Singh, Ram Nevatia, and Chang Huang, "Efficient inference with multiple heterogeneous part detectors for human pose estimation," in *ECCV*, 2010.