Informative Motion Extractor for Action Recognition with Kernel Feature Alignment

Taketoshi Mori Masamichi Shimosaka Tatsuya Harada Tomomasa Sato
Graduate School of Information Science and Technology
The University of Tokyo
Tokyo, Japan
Email: {*tmori, simosaka, harada, tomo* }@ *ics.t.u-tokyo.ac.jp*

***Abstract*— This paper proposes a novel algorithm for extracting informative motion features in daily life action recognition based on Support Vector Machine(SVM). The main advantage of the proposed method is not only to extract remarkable motion features which fit into human intuition, but also to improve the performance of the recognition system. Concretely speaking, the main properties of the proposed method are 1)optimizing kernel parameters so as to minimize its generalization error, 2)extracting remarkable motion features in response to the sensitivity of the kernel function. Experimental result shows that the proposed algorithm improves the accuracy of the recognition system and enables human to identify informative motion features intuitively.**

I. Introduction

Intelligent computational systems like robots are expected to support humans in everyday tasks and activities in the near future. To achieve assistance in a wide range of areas, it will be important for that kind of systems to be able to communicate effectively with human. Recognizing human actions has potential to contribute to this ability and many other applications, such as human computer interaction and search engine for multi media databases.

It is proper to divide the process of action recognition into the following two phases. The former is to acquire time series of three-dimensional body motion structurally from some instruments, such as multiple images sensor systems and infrared motion capture systems. The latter is to symbolize these kinds of motion to action names. This hypothesis has roots in MLD[1]. As the former phase, a state of the art technique of marker-less human motion tracking such as Deutshcer's[2] is actively developed in recent years. But it is rare that these kinds of systems robustly work in real-time. Our research focuses on the latter phase.

As the latter phase(symbolizing recognition), Shimosaka et al.[3] developed a recognition system for such a daily life action as walking and sitting, whose performance is optimized by Support Vector Machine(SVM). The main remarkable characteristics of the system is to utilize expressions of action by human knowledge.

The approach produces the following merits. At first, this enables the system to improve recognition performance easily and to be built intuitively. Secondly, this approach aids for designer of the system to detect remarkable motion features in motion candidates.

But at the same time, humans' expression-based approach causes some critical problems as follows. Firstly, it is difficult to generate expressions in some actions. Secondly, there is difficulty of selecting relevant motion feature from the generated expressions in some action. Finally, it is difficult to determine which motion features is important quantitatively.

Thus, this paper proposes an algorithm that automatically extracts informative motion features corresponding to the target action quantitatively from reference motion data. The proposed method differs from other feature extractor in the following points. One significant point in the aspect of performance is that the proposed method extracts informative motion in the single criterion of classification and extraction, meanwhile common classifier with some feature extractor such as Fisher discriminant analysis or principal component analysis uses different criterion on feature extraction and classification. The significant point in the aspect of feature extraction using kernel[4] is that the proposed method is well-suited to "knowledge discovery", meanwhile other kernelized feature extractor is not designed for "knowledge discovery".

In section 2, an action recognition system as the basis of the proposed algorithm will be described. The qualitative and detailed description of our knowledge discovery approach, which utilizes kernel parameters optimization, will be explained in section 3 and 4, respectively. The performance evaluation of the proposed algorithm will be noted in section 5.

II. Daily Life Action Recognition System: HARS

As for the basis of the proposed algorithm, we constructed a SVM-based action recognition system. It takes over the characteristics of Shimosaka et al.'s system[3] and Mori et al.'s[5] mentioned below.

First feature is *Simultaneous Recognition*. This is because human can recognize multiple action at the same time. For example, human can readily recognize someone waving his or her hand while walking as "waving one's hand and walking".

Second feature is *Unclarity in Recognition*. This is because human cannot always give absolute decision whether some action really occurs or not when watching someone acting. For instance, decision whether lying or not made by human may contain unclarity on observing someone

getting up. Consequently, our system is designed to be able to output multiple action name labels at the same time and to output not only decisive result, but unclarity result.

In Mori et al.'s system[5], an approach, *Utilizing Expressions of Action by Human*, is adopted. This has arisen from feature selection problem, because such remarkable features for recognizing some action as motion and pose of body region have wide variation. For example, forward motion of hips could be one of the features for walking, meanwhile the direction of head is considered as irrelevant. Because human can easily express an action by representing the motion or the pose of body parts, a designer of the system selects input motion with aid of these expressions. Because this manual fashion approach invokes some critical problem noted above, the necessity of this research proposing automatic extractor becomes clear.

In order to make us evaluate performance of a recognition system easily, the correctness of the recognition result by humans is defined as follows. On observing someone acting, even if he or she is performing several actions simultaneously, reference data of recognition result is generated in synchronized with input as the discrimination of one action by paying attention to this action only. For instance, if the target action is assigned as "walking", human pays attention to "walking". Thus, the reference is equal to the result whether someone is walking, even if he or she actually "walking" and "standing" at the same time.

A. Input and Output

The input of the system is time series of human motion. Concretely speaking, our system utilizes articulated human body motion whose three-dimensional configuration is recovered. The output of our system contains some action names in synchronized with a frame of the input motion.

B. Configuration of HARS

Figure 1 shows the processing flow and the configuration of the system. In order to realize the simultaneous recognition, the system contains multiple recognition processes, each of which is assigned to the recognizer for one action. This primitive process is called as "Action Element Recognizer(AER)". One AER runs in parallel with the others. The system collects the results of all AERs, and outputs the results of each recognition process per frame. An AER which recognizes walking discriminates whether someone is walking.

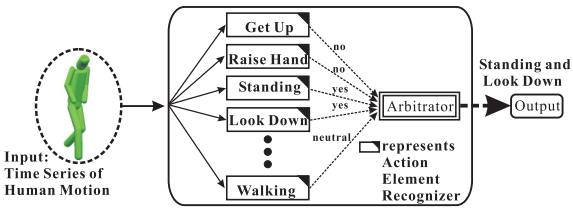

Fig. 1. Configuration of Recognition System

C. Kernel based Action Element Recognizer

Time series of human motion is utilized as input of each AER. An AER outputs result of classification whether the assigned action occurs or not per frame in synchronized with the input motion. The output of AER consists of multiple classes, which represent not only decisive but also inexplicit result. Concretely speaking, the number of the category is three. One category is named as "yes" which represents that the assigned action clearly occurs. Another is named as "no" which represents that the opposite meaning of "yes". The last one represents the unclarity of recognition result called as "neutral".

The AER contains two binary classifiers and outputs integrated result of the two binary values. Concretely speaking, the one binary classifier judges whether "yes" or not-"yes", the other judges whether "no" or not-"no". The configuration of the AER is shown in Figure 2. The reason why we have adopted this composition is that "neutral" category rarely happens in some actions. For example, human can explicitly discriminate motion as sitting down when watching someone standing then sitting.

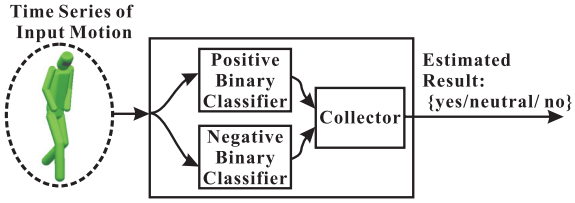

Fig. 2. Configuration of Action Element Recognizer

Binary Kernel Classifier: A kernel classifier is introduced as the binary classifier in the AER. We denote by $\boldsymbol{x}$ the time series of input motion. $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^{l}$ are the input-output pairs in total l frames, where $\boldsymbol{x}_i$ represents i th frame sample motion and its corresponding reference binary(e.g. "yes" or not-"yes") signal by y_i. We can write by α_i the co-efficiencies whose value is proportional to importance of the templates. Similarity value between the input motion and one template motion is represented by Kernel $K(\boldsymbol{x}_i, \boldsymbol{x})$. The mapping between the input and the output of the binary classifier in the AER f can be written as

$$f(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right)$$

where b depicts offset and the function $\mathrm{sgn}(\cdot)$ is a step function where the relation of input-output is represented as

$$\mathrm{sgn}(t) = \begin{cases} +1 & \text{if} \quad t > 0 \\ -1 & \text{if} \quad \text{otherwise} \end{cases} .$$

Learning process in the binary classifier of AER tunes the co-efficiencies($\boldsymbol{\alpha}, b$) from the training data. SVM[6] is utilized as learner in it. SVM is one of the honored learning algorithm in the view of regularization, model selection and requirements of the computation resource.

Kernel as Product of Kernels per Gazed Motion: The kernel value in the AER binary classifier is as the product of all the kernel values corresponding to the similarity in each gazed motion. When the number of the gazed motion in the target action is d, the kernel value in the target action $K(\cdot,\cdot)$ can be written as

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\prod_{j=1}^{d}K_j(\boldsymbol{\varphi}_j(\boldsymbol{x}^{(j)}),\boldsymbol{\varphi}_j(\boldsymbol{x}_i^{(j)}))$$

where $\boldsymbol{x}^{(j)}$ denotes the selected input motion in the j th gazed motion, $\boldsymbol{\varphi}_j(\cdot)$ represents the converter from the selected input motion to the input feature, and the kernel value which corresponds to the similarities in j th gazed motion represented by $K_j(\cdot,\cdot)$. In this paper, the gazed motion means the candidate of the remarkable motion for recognizing the target action. In general, Radial Basis Function is utilized as the kernel for gazed motion, thus the final form of the kernel is represented as

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\sum_{j=1}^{d}\frac{\left|\boldsymbol{\varphi}_j(\boldsymbol{x}^{(j)})-\boldsymbol{\varphi}_j(\boldsymbol{x}_i^{(j)})\right|^2}{\sigma_j^2}\right). \quad (1)$$

III. Remarkable Motion Extractor based on Kernel Parameters

It is the fundamental premise that the kernel types and their parameters are priori given in the learning process of any kernel classifiers, and the performance is surely dependent on the kernel types and their parameters. However SVM achieves more honor than other classical learning algorithms, the performance fails to acquire high accuracy when some kernel types and its parameters are set. Thus the functionality that adjusts the kernel feature space must be needed in order to build more suitable SVM.

It is natural to think that sensitivity of the kernel value with respect to change of input should be large if it is relevant for recognition. On the other hand, the smaller sensitivity of the kernel value might be desired in the case of no importance. As for Mahalanobis kernel utilized in our recognition method, the remarkable input feature requires smaller variance σ in relevant input feature than in irrelevant.

In opposite way, if one wants to know which input feature is relevant, it is somewhat appropriate that he or she judges which input motion is remarkable by observing the kernel sensitivity. As for Mahalanobis kernel, the inverse of the variance σ in Eq.(1) helps the designer of the recognition system to detect which motion is informative. This is because these parameters affect the sensitivity of the kernel values.

This paper utilizes the kernel parameters optimization which adjusts kernel sensitivity in order to extract remarkable motion features and to optimize its performance at the same time. In the case of our system, the proposed algorithm adjusts the variances in Eq.(1) and trade off positive number in SVM.

IV. Kernel Parameters Optimization

In this paper, the generalization error is utilized as the indicator of the optimization. Therefore, the optimized kernel parameters $\boldsymbol{\theta}^*$ is defined as $\boldsymbol{\theta}^*=\arg\min_{\boldsymbol{\theta}}T(K_{\boldsymbol{\theta}})$, where $\boldsymbol{\theta}\in\mathbb{R}^{d_k}$ denotes the kernel parameters, $T(K_{\boldsymbol{\theta}})$ depicts the generalization error of the SVM.

In order to optimize kernel parameters easily and robustly, an effective search in kernel parameters space must be considered. In this paper, the gradient descent algorithm is utilized. General outline of the kernel parameters optimization algorithm is listed as Table I.

TABLE I
Kernel Parameters Optimization based on Gradient

1.	Initialize $\boldsymbol{\theta}$ with some value and iteration number i as 0
2.	Learning by SVM with $K_{\boldsymbol{\theta}}$, finding co-efficiencies.
3.	Calculating generalization error T and its derivative.
4.	Update the kernel parameter $\boldsymbol{\theta}$ such that $T(K_{\boldsymbol{\theta}})$ is minimized as $\Delta\boldsymbol{\theta}=-\epsilon\partial T/\partial\boldsymbol{\theta}$, $\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}+\Delta\boldsymbol{\theta}$, $i\leftarrow i+1$. $(\epsilon>0)$
5.	In the case that $\|\Delta\boldsymbol{\theta}\|$ is less than some positive constant or i is larger than some iteration times, then terminate, otherwise return to **2.** .

There are several good estimators for the performance of SVM. In this research, the *Span* technique proposed by Chapelle et al.[7] is adopted. This is because the gradient descent technique requires the derivative function of generalization error by kernel parameters and it can be explicitly written if estimator is based on *Span*. The property of generalization error with *Span* has close relationship with the Leave-One-Out cross-validation error(LOO). The computational cost of LOO is too high but accuracy gets excellent quality. In contrast, the *Span* technique requires less computational resource than the case of LOO.

A. Span Bound of Generalization Error

After learning by SVM, the span corresponding to the p th support vector by the variable S_p can be defined as the distance between the $\boldsymbol{\phi}(\boldsymbol{x}_p)$ and linear combination Λ_p by all the support vector except the p th support vector in feature space $\boldsymbol{\phi}(\cdot)$ as

$$\begin{aligned}\Lambda_p &= \left\{\sum_{i\neq p,\alpha_p>0}\lambda_i\boldsymbol{\phi}(\boldsymbol{x}_i),\ \sum_{i\neq p,\alpha_p>0}\lambda_i=1\right\}\\ S_p^2 &= \min_{\boldsymbol{\phi}(\boldsymbol{x})\in\Lambda_p}||\boldsymbol{\phi}(\boldsymbol{x}_p)-\boldsymbol{\phi}(\boldsymbol{x})||^2=\frac{1}{(\tilde{K}_{sv}^{-1})_{pp}}.\end{aligned}$$

where function $\boldsymbol{\phi}:\mathcal{X}\rightarrow\mathcal{F}$ ($\mathcal{F}$ represents some feature space) satisfies $\boldsymbol{\phi}(\boldsymbol{u})^t\boldsymbol{\phi}(\boldsymbol{v})=K(\boldsymbol{u},\boldsymbol{v}),(\boldsymbol{u},\boldsymbol{v}\in\mathcal{X})$ and $\tilde{K}_{sv}$ corresponds to extended Gram Matrix of all the support vectors. The upper bound of the generalization error based on the *Span* is defined as

$$T_u=\frac{1}{l}\sum_{p=1}^{l}\Psi(\alpha_pS_p^2-1)$$

where $\boldsymbol{\alpha}$ denotes the co-efficiency obtained by SVM, Ψ depicts step function to penalize.

Because the optimization process requires the derivative by the kernel parameters, the gradient of Ψ, $\boldsymbol{\alpha}$, S_p^2 must be calculated. The probabilistic approach by approximating the step function Ψ with sigmoid function[8] is adopted. In this case, Ψ is approximated as a sigmoid function as

$$\Psi(t) \equiv \frac{1}{1+\exp(-At+B)}, \ A > 0, \ B \geq 0.$$

The parameters A and B in the sigmoid function can be estimated by minimizing Kullback-Leibler Divergence. The gradient of the co-efficiency $\boldsymbol{\alpha}$ can be calculated because the relation between output and input of the SVM can be written only by support vectors. The computation for the derivative of S_p^2 can be calculated by utilizing Woodbury Theorem[9], and finally this is derived as

$$\frac{\partial S_p^2}{\partial \theta_q} = S_p^4 \left(\tilde{K}_{sv}^{-1} \frac{\partial \tilde{K}_{sv}}{\partial \theta_q} \tilde{K}_{sv}^{-1} \right)_{pp}$$

where θ_q represents the q th kernel parameter.

V. Experiments

A. Target Action and Motion Candidates

In order to evaluate the performance of the proposed method, 18 action names, such as "Standing", "Folding Arms" are selected and ICS Action Database[10] are used. The specification of the motion used in the experiments is listed as Table II. This is the collections of motion data with reference action name labels. BVH, the format of motion data, is a de-facto standard in computer graphics by the Biovision Corporation. A BVH file contains the structure of a human as a linked joint model(figure) and the motion of the figure per frame. The skeletal configuration of the BVH used in the experiment is shown in Figure 3.

TABLE II
Specification of ICS Action Database

D.O.F.	36(11 Articulation)
Actor	a male in 20s
Format	Biovision BVH and its label
Num. of Files	125 (Avg. 3.2[sec.])

125 BVH files and 2,250 reference files are used in our experiments. The label in the database representing humans' judgment has three kinds of values (yes, neutral, no) in every frame per one action name. Figure 3 depicts all 18 action names and snapshots of them. Although this action database contains labels of 25 action names, 18 names are selected. This is because the time when the unselected action names, such as "Walking" and "Sit Down" occur is short.

As for candidates of remarkable motion feature, the human motion feature utilized in Shimosaka et al.'s system[3] are selected. The entire selected motion features are listed as Table III. The bracket after candidates(right side of the table) represents the duplication of candidates, each of whose span is different.

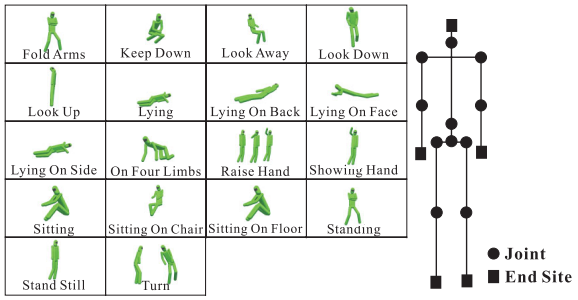

Fig. 3. This figure shows thumbnails of Target Actions(in left side) and skeletal configuration of BVH used in this experiment(in right side). The body contains 11 joints each of which has 3 degrees of freedom.

TABLE III
Enumeration of Given Motion Information

1	Relative horizontal position of right hand to left	**2**	Sum of distance between hands and body
3	Mean speed of hands	**4**	Bentness of Hips
5	Mean speed of hips(1)	**6**	Height of hips
7	Upper direction of head from hips	**8**	Distance between hips and foots
9	Upper direction of hips from foots	**10**	Horizontal orientation of head from hips
11	Upper orientation of head from hips	**12**	Upper orientation of head from ground
13	Mean height of head(1)	**14**	Height of hips
15	Upper orientation of hips	**16**	Horizontal orientation of hips
17	Upper direction of head from left hand	**18**	Upper direction of head from right hand
19	Upper direction of hips from left knee	**20**	Upper direction of hips from right knee
21	Height of left hand	**22**	Height of right hand
23	Relative height of left hand from hips	**24**	Relative height of right hand from hips
25	Mean upper velocity of left hand	**26**	Mean upper velocity of right hand
27	Highest relative height of hands from head	**28**	Mean height of head(2)
29	Mean speed of hips(2)	**30**	Mean speed of left foot
31	Mean speed of right foot	**32**	Speed of Rotation of hips in vertical axis

B. Given Parameters and Condition

As the initial parameters of this experiment, the parameters of Mahalanobis kernel and trade-off number in SVM C are given as $\sigma_j = 1.5\sqrt{f_d}$ $(1 \leq j \leq f_d)$, $C = 5\sqrt{f_d}$ in all 18 action names, where f_d denotes the dimensionality of the kernel input space (i.e. 32). The Maximum iteration times of the gradient descent procedure is set as 30. As the updating rate ϵ of the second procedure in the table I is set as $\epsilon_\sigma = 0.05$ for kernel parameters and $\epsilon_C = 0.1\epsilon_\sigma$ for penalty term.

C. Result of Applying Proposed Method

Average of Accurate Rate: Figure 4 shows that error rate the before and after the kernel parameters optimization in the case that all the 32 motion features are utilized. As a whole, the accuracy of the recognition in each target action achieves 80[%]. Especially, the accurate rate in 14/18 action names is larger than than 90[%].

In almost all the target actions, the accuracy gained by the optimized kernel parameters is better than the case of the initial kernel parameters. Unfortunately, the accurate rate reported in [3] is better than the proposed method. This result shows that the strategy which utilizes expression of action works good only if expressions can be generated easily.

Furthermore, the target action “Stand Still” gains less accurate after the kernel parameters optimization. It is found that the variance corresponds to velocity of hips is much smaller than the others. It seems that this causes the decline of the accurate rate. Precise analysis for not only “Stand still” but also “Turn” must be one of the future work of our research.

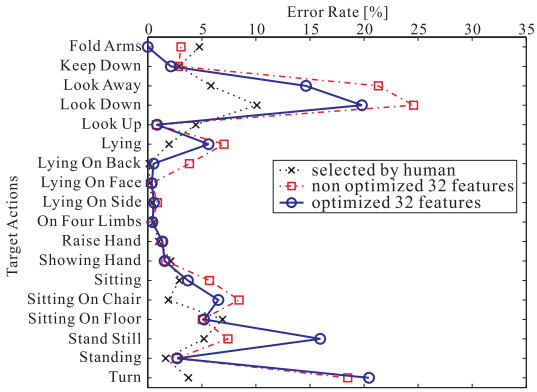

Fig. 4. This figure illustrates error rate before and after kernel parameters optimization in each action. Error rate in the case of the initial given kernel parameters are represented by the square points. Circle points depict the case of the optimized kernel parameters and cross points represent the score reported in [3].

Ratio of the gained kernel parameters: Figure 5 and 6 shows the relative ratio of the kernel parameters after the procedure of the proposed algorithm in the case of “Standing” and “Raise Hand”, respectively. In each figure, the number on the horizontal axis corresponds to the number in Table III. The vertical axis shows the normalized inverse of the variances whose maximum value is 1. In the scheme of the proposed detection method, the larger value in the vertical indicates more relevant motion feature, because smaller kernel parameter σ makes kernel more sensitive.

In the case of “Standing”, inverse of the variance corresponds to bentness of hips is largest. Next, the orientation of the upper body is detected, and horizontal posture of hips is detected as the third relevant motion feature. This result fits into human intuition.

As for “Raise Hand”, the proposed algorithm detects velocity in the upper orientation of right hand only. It is found that there is no “Raise Hand” motion where left hand moving upward is observed in the training database. Thus, no proper priori idea like symmetry of action gives improper result, but this result shows that the relation between the gained ratio of the kernel parameters and motion in the training data seems to be natural.

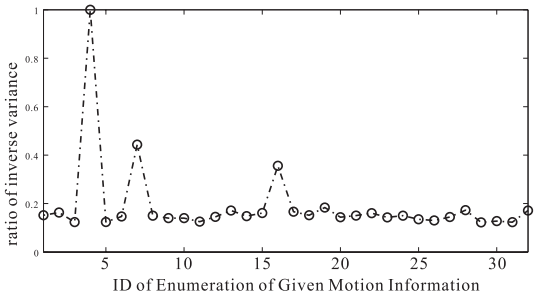

Fig. 5. Normalized inverse variance after optimization in “Standing” are shown. ID 4 in Given Motion Information represents “Bentness of hips”.

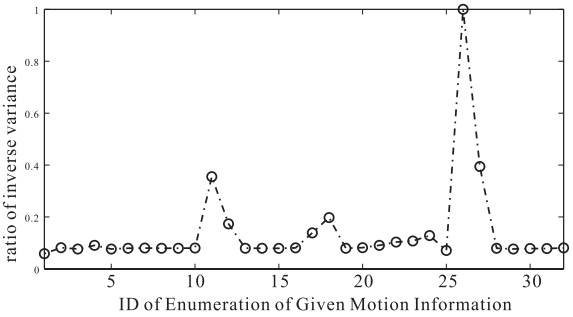

Fig. 6. Normalized inverse variances after optimization in “Raise Hand ” are shown. ID 26 in Given Motion Information represents “Mean upper velocity of right hand”

D. Validation of the Detection Scheme

As a validation of the detection scheme that kernel sensitivity responses to remarkableness of motion features, the performance obtained by the selected motion feature judged by inverse of the variances is evaluated. Concretely speaking, some human motion features, whose corresponding inverse of the variance is large(large value in the vertical axis of the Figure 5 and 6), is selected from all the candidates. In this experiment, the number of the selected motion features is set to 7. The reason why we set the number as 7 is not clear, but it is thought that 7 selected features contain adequate relevant motion information in each action because the maximum dimension of features reported in [3], [5] is 7. As this experimental condition, the variances are set to equal after the selection.

Figure 7 shows the error rate in each target action gained by the selected 7 features and optimized 32 features. In all the target action except “Look Down”, the accurate rate acquires 90 [%]. As for “Look Down”, because the accurate rate in the case of the optimized 32 features fails to acquire 90 [%], the selected 7 features from these features also seem to have trouble to recognize. The result noted above experimentally implies that the detection scheme based on the kernel sensitivity is valid.

As another validation of the detection scheme, the performance obtained via motion features by removing large inverse variance is evaluated. It can be said that the previous validation evaluates the affection of removing

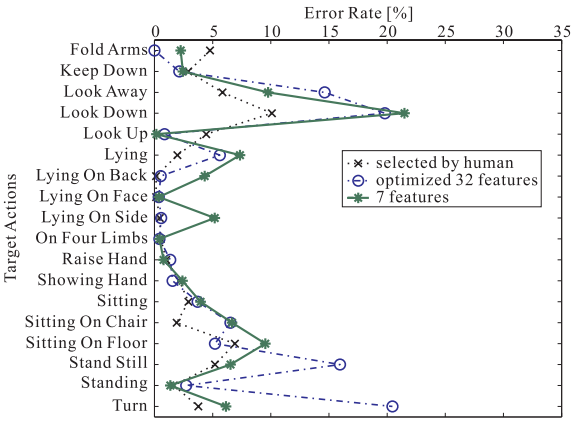

Fig. 7. This figure shows the error rate gained by selected 7 features which is thought to be remarkable. The performance gained by the optimized 32 features in the previous experiment is also shown.

irrelevant motion features. On the other hand, this experiments evaluate the importance of the motion feature whose kernel parameter(invariance variance) is large. Especially, the number of the removed feature is 1. After the removal, the variances are set as equal. Furthermore, the kernel parameters optimization is executed for such 31 dimensions.

The performance score in each condition is shown in Figure 8. The result shows that the performance of the all the targets except “Keep Down” and “Sitting On Chair”, “Standing” obtained by the optimized kernel parameters from 31 dimensions is worse than the others. This result implies that the removed largest inverse of the variance motion feature is fatal to recognize, thus the kernel parameters optimization via 31 dimension fails to acquire generality but over-fitting. This result indicates the proposed detection scheme based on inverse of the variances(generally speaking, the kernel sensitivity) seems to be valid.

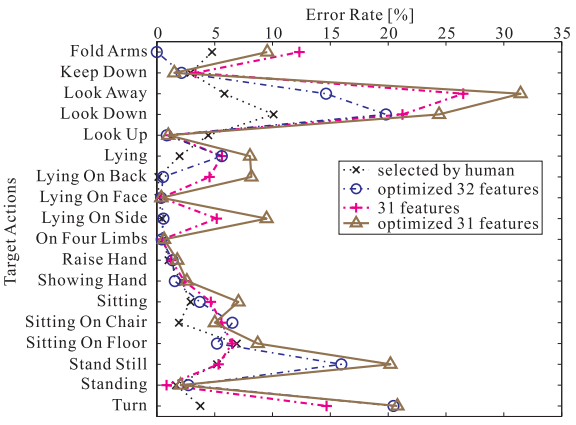

Fig. 8. This figure illustrates the error rate obtained via 31 features by removing 1 feature whose inverse variance is largest. Both performance before and after kernel parameters optimization are shown. The performance gained by 32 motion whose kernel parameters are optimized is also shown.

VI. Conclusion

This paper proposes an algorithm for extracting remarkable motion features from candidates of motion features in human daily life action recognition based on kernel classifier. This algorithm is based on kernel parameters optimization as minimization of generalization error. In this paper, *Span* based generalization error which can be calculated effectively and has close relationship to Leave-One-Out cross-validation error is utilized. This paper adopts gradient information because search method for kernel parameters space must be effective.

The experimental result for performance evaluation shows that the accuracy of recognition achieves high enough, in addition, the performance after the optimization of the kernel parameters is better than the case of the initial settings. It is also proved that the relative importance values which corresponds to inverse of the variances(the kernel value in Mahalanobis kernel) fits into human intuition. The other experiment proves that our detection scheme is valid in the point that the motion features corresponding to some large sensitivity of kernel functions is critical to recognize.

We have plan to apply the proposed method as the following way. In recognition system using marker based or wearable based motion capturing system, the proposed method helps the designer to decide the minimum essential sensors to recognize the target actions, because these kinds of sensors burden the actor.

In the future work, systematic method for listing motion candidate will be explored, because this paper lists the candidates of the motion features utilized in [3]. Next, analysis between density in input space and its sensitivity of kernel feature space will be considered, because the proposed algorithm initializes the parameters which satisfy range of kernel value is same in each feature.

References

[1] G. Johansson. “Visual Perception of Biological Motion and a Model for its Analysis”. *Perception and Psychophysics*, pp. 201–211, 1973.

[2] J Deutscher et al. “Articulated Body Motion Capture by Annealed Particle Filtering”. In *Proceedings of the 2000 IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2126–2134, 2000.

[3] M. Shimosaka et al. “Recognition of Human Daily Life Action and Its Performance Adjustment based on Support Vector Learning”. In *CD-ROM of Third IEEE International Conference on Humanoid Robots*, 2003.

[4] B. Schölkopf et al. *Learning with Kernels*. MIT Press, 2002.

[5] T. Mori et al. “Human-like Action Recognition System Using Features Extracted by Human”. In *Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1214–1220, 2002.

[6] V. Vapnik. *The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science)*. Springer Verlag, 1995.

[7] O. Chapelle et al. “Model Selection for Support Vector Machines”. In *Advances in Neural Information Processing Systems 12*, pp. 230–236. MIT Press, 2000.

[8] J. Platt. “Probabilities for SV Machines”. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

[9] H. Lütkepohl. *Handbook of Matrices*. Wiley & Sons, 1996.

[10] ICS Action Database. http://www.ics.t.u-tokyo.ac.jp/action/, 2003.