Replies to comments (All included)

We happily appreciate your great and precious time during the review of our paper and the detailed comments. With respect to the suggestions, we have deeply revised the manuscript. Your comments and questions are printed in the normal font and our replies are in blue print.

Review1.

English language and style
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	(x)	( )	( )	( )
Is the research design appropriate?	(x)	( )	( )	( )
Are the methods adequately described?	( )	(x)	( )	( )
Are the results clearly presented?	()	(x)	( )	( )
Are the conclusions supported by the results?	( )	(x)	( )	( )

Comments and Suggestions for Authors

The paper is interesting and contains good variety of numerical results. The theoretical part, however, needs to be improved.

In Section 2, the style of Eq (4) differs from the style of the formulas above and the operator \Sigma is not defined. On line 83 the formula for the number of nodes for the OVO method is probably wrong, as later it seems that the formula k(k-1)/2 is used. Lines 92-93 contradict with the definitions of OVA and Ordinal, as true and false for the two methods were defined to be opposite. Formula (5) looks strange - in the left-hand-side the summation is until infinity, while in the right-hand-side - it is only up to n. Please, check if this is indeed the correct formula.

Sigma in Eq.4 should have been covariance, which appears to be a mistake during editing. It is changed same with our intention.

$d_{mahalanobis} = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$, where S is the covariance

In lines 92-93, we found that the definitions of OVA and Ordinal can be ambiguous. In case of OVA, one SVM is used per classification target class, classifying the class as 1 (True) and the rest as False (-1). In the case of Ordinal, the number of nodes needed was reduced by 1 using all classification results of nodes. In this case, the target class has a different bit configuration depending on sorting order. Our previous manuscript overlooked this difference, so we changed the description to clarify the definition, and rewritten as:

Figure 1 shows the bit composition of the OVA and ordinal methods for a 7-class problem. In the case of OVA, each bin (b) corresponds to an SVM classifier which returns a value of 1 for the target class as true and returns -1 for other classes as false. In the case of Ordinal, a combination of returns by b1~b6 makes decision to classifying classes. OVA consists of k SVM classifiers, and ordinal reduces the number of bins by 1 through a modified sorting order.

Parameter t had to be increased from 0 to n, not to infinity, so we changed INF to n for our intention. Eq.5 is rewritten as:

$$\sum_{t=0}^{n} \gamma^t R_{\pi(s_t)}(s_t, s_{t+1}) = E[R(s_0, s_1) + \ldots + \gamma^n R(s_{n-1}, s_n)]$$

In both Section 2 and 3 there are plenty of unnecessary repetitions in the formulas, the definitions, etc. The content of those sections seem to be classical and not very original, so my suggestion is the authors to try to shorten the exposition and avoid repetitions (e.g., Eq (13) and Eq (27)). There are no punctuation marks after the majority of the highlighted formulas. In Eq(18) it is better to include the left-hand-side N_b=...

Repeated Eq 13 and Eq 27 are mistakes during editing. The intention of them was the Q function of Q-learning can be approximated to the expression of accuracy (A) by a relationship of Eq 26, and we changed Eq.28 into bellowing which same with our intention.

$$A(s_t, a_t) \leftarrow (1-\alpha)\, A(s_t, a_t) + \alpha r_t + \gamma \max_a A(s_{t+1}, a)$$

Morefore, some changes are given to Eq.14~19 for clarifying definition, and 'N_b =' is also added in Eq.18

$$N_b := N_{center} + 4N_{overlap} \qquad (15)$$

$$N_{center} := N_{gradient} \times c^2 \qquad (16)$$

$$N_{overlap} := N_{gradient} \times (0.5c - 1)(1.5c - 1) \qquad (17)$$

$$N_b = N_{gradient} \times \left(c^2 + 4(0.5c - 1)(1.5c - 1)\right) \qquad (18)$$

$$H^* = (H_1^T, H_2^T, \ldots, H_k^T)^T \qquad (19)$$

On line 310 it is written that OVA outperforms OVO in Table 2, but the documented results in the table show the opposite. Formula (28) looks strange, as after simplification it can be written in the form \alpha=-c\sum_{k=0}^n A(s_k)/n. Is this correct? In the caption of

Figure 17, the image legends are switched - Grid Map is on the right, while basic HOG is on the left.

It was a mistake to describe OVA on line 310 as best, so it changed to OVO, and rewritten as:

As shown in Table 2, OVO resulted in the highest accuracy and the lowest classification time.

The purpose of Eq.28 is to adjust the learning-rate factor based on decreasing variance of classification accuracy by dividing the Grid Map as the policy updating. When the total number of divisions is n, variances of all possible divisions in Grid Maps are added and divide by n to adjust the learning rate. Previous equation in manuscripts can be differed from our intention. Therefore, we changed Eq.28 as definition that described above, and rewritten as:

$$\alpha := \frac{c}{n}\sum_{k=1}^{n}\left(\sum_{\mathrm{B_i} \in G(s_k)} \{A(s'_{B_i}) - A(s_{k-1})\}^2\right)$$

We corrected descriptions that makes differ with our intention in Figure 17.

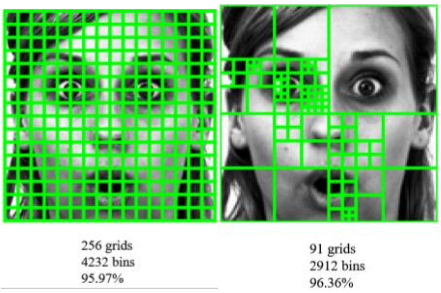


Figure 17. Feature Extractor cell distribution of basic HOG (left) and Grid Map (right).

Review2.

English language and style
( ) Extensive editing of English language and style required
(x) Moderate English changes required
( ) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	( )	(x)	( )	( )
Is the research design appropriate?	(x)	( )	( )	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	(x)	( )	( )	( )
Are the conclusions supported by the results?	(x)	( )	( )	( )

Comments and Suggestions for Authors

A modified SVM called Grid Map is proposed for facial expression recognition. Regional Weighting is proposed in which weights are assigned during feature extraction rather than applying weights at the training stage of classifiers. Experiments are performed on public databases of facial expressions. Few methods are compared for classification and it shown that the proposed approach using ECOC+SVM for classification is comparable to existing approaches using fewer features and less computational resources. However, the computational cost of creating the adaptive grid map needs to be considered. It needs to be clarified if the feature extraction process will take longer for the proposed approach.

That content was written in section 4.3 Result of Features Reduction of manuscript. Previous manuscript has unclear description of Classifying a image (ms). This experiment surely includes the computational cost of image crop process and others for creating the adaptive grid map. Now, explanation has been added to clarify this section and corresponds to 395-397.

Table 4 shows computational cost of basic and proposed method. Each total time (s) and classifying an image (ms) include all computation for classifications with a single thread of an Intel Core i7-8750H 2.3GHz processor. Proposed method shows less computational cost even has more processes.

Review3.

English language and style
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	( )	( )	(x)	( )
Is the research design appropriate?	(x)	( )	( )	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	()	(x)	( )	( )
Are the conclusions supported by the results?	( )	( )	(x)	( )

Comments and Suggestions for Authors

The authors presented a new approach to recognize facial expression using Q-learning. The authors proposed a modified support vector machine with a combination of reinforcement learning. The authors proposed a different approach of feature extraction using Regional Weighting which contains more information and helps to increase recognition accuracy. However, I have some concerns regarding the structure of the paper and the content of it.

Usually, the abstract must reflect all the content of the article in a very concise form and must contain the proposed methodology, results, and conclusion from the results. The authors did not include the results obtain

(the accuracy in numerical value). The authors most include at least a sentence with the results obtained.

Numerical results are added in the Abstract. We had to explain the research within 200 words, so I think it was our mistake to exclude the numerical results. We have added sentences about numerical results, except for relatively unnecessary explanations. The explanation that are excluded are:

~~, and it was then used as the reward value in Q-learning for optimizing Grid Map using 10-fold cross-validation~~

And the sentences that are included are:

The proposed method is formulated into a decision process and is solved using Q-learning. Proposed method shown classification accuracy of 7-emothions 96.36% in 4 database and 98.47% in ck +. Comparing its computational cost to basic method with similar accuracy, 68.81% features and 66.33% processing time are required.

Introduction section: the authors fail to include more relevant articles that motivate you to carry out this research.

Let me elaborate on this issue:

The basic purpose of the research study is to propose a method to find better results than previous studies. For that, authors need to mention most related previous studies, point out their limitations, and

propose your methods to improve the results. This must link to your methodology, results, and discussion.

It is suggested to elaborate more on the introduction with some more relevant studies, their limitations, and your proposal.

The main motivation of our research is to improve the limitations of previous facial components weighting methods with reinforcement learning. Existing Introduction section has seemed to lack the analysis of limitation and classification accuracy of facial landmarks (FLs) and facial action units (AUs) which is existing facial component weighing methods. Therefore, we added explanations about analysis of previous methods as limitations of that study.

The overall flow of our research suggests and validates a Grid Map that replaces the regional weights of facial elements such as facial landmarks (FLs) [7] and facial action units (AUs) [8]. These methods of weighting facial elements have efficiently improved classification accuracy through consideration for importance. However, these studies have a limitation that a human forced definition of clustering rules for each element is required. To overcome these two limitations, our study proposes a adaptively defines the optimal feature extraction cluster according to reward maximization in reinforcement learning. Increasing classification accuracy with efficient feature extraction affects rewards during training. In addition, reinforcement learning has not been attempted for the improvement of facial element weighting efficiency. The application of dynamic programming is considered difficult in this context because the configurations of weighting models cannot be changed by conventional methods. Therefore, our study also includes a validation of whether classification accuracy is affected by the feature detail differences of facial element regions, and we propose a weighting model called Grid Map, which considers differences of details, and we update its values using reinforcement learning. This Grid Map contains a regional distribution of bounding grids and is optimized for maximum accuracy by combining a multiclass-SVM classifier using HOG (histogram of gradients [9])-ECOC (error-correcting output codes [10]) classification as a combining method for reinforcement learning.

Section Two and Three:

The authors presented a related theory that includes a detail illustration of mathematical equations.

Section Four (Experiments):

Suggested to present used databases in more compact form eg: Table.

Figures 9 and 10, or especially 10, Neither readers have the interest to see these images nor they can better present only a few images or remove the figure.

It was replaced with a figure that illustrates that the face area is regularly cropped.
Following figure is replaced one.

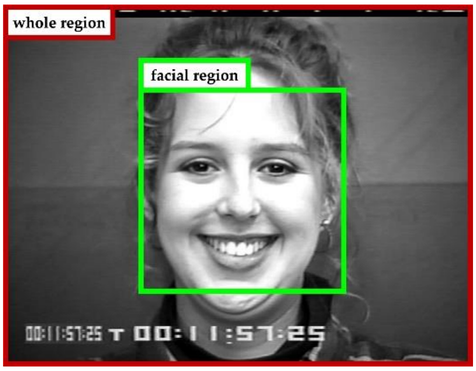


Figure 9. A frame in S052.004 sequence of CK+ Database

Figure 9 is replaced to explain uncessary region(out of green box) and facial region(inner of green box). Purpose of database regularization is normalizing position of facial components(eyes, nose, mouth etc.). So we made a new figure that focused on facial region and its bounding.

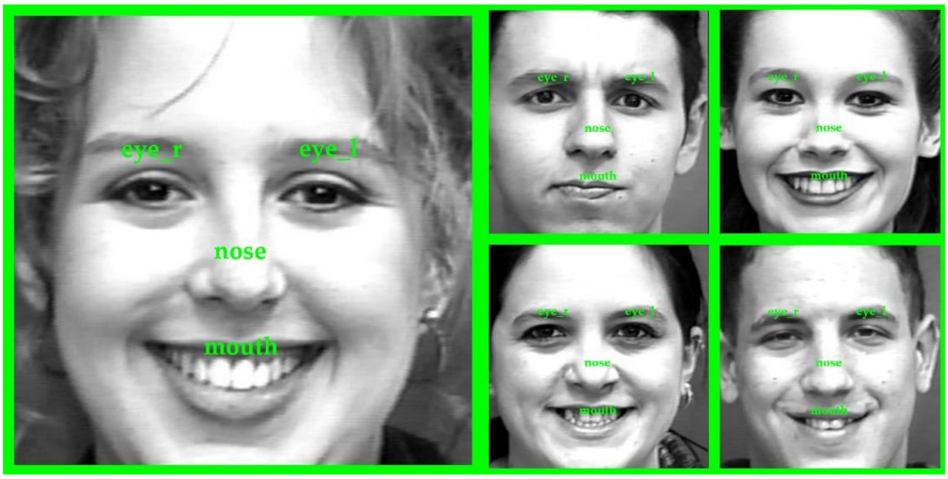


Figure 10. 5 images of modified CK+ database acquired through the database regularization with same cascade rule

Figure 10 is replaced to illustrate regularized database caused by using same cascade. Previous figure was hard to see because contating lots of small images. So we selected only 5 images as shown in Figure 10 with each nested cascade(eyes, nose, mouth) in a facial cascade.

Recheck "The optimal cell size was found to be 10 and 14". Section 4.1. Line-303.

Cell size 14 was also 97.44% accurate. However, since the precision from the third decimal point appeared to be the same, it was clarified to be 10.

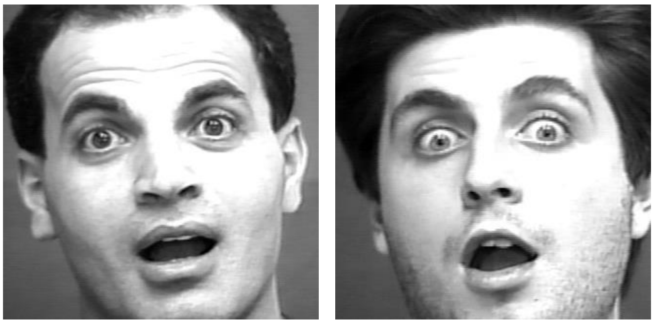
As shown in Table 1, the HOG-ECOC method shows the maximum accuracy and classification speed. The optimal cell size was found to be 10.

Move this text to the discussion section page from section 4.3 the last paragraph. "However, further validation is required in future research. Additional experimental 414 results are also needed regarding the application of our proposed reinforcement learning to a study 415 of classification accuracy improvement using a modified binary tree SVM to classify Neutral to 416 other 6 expressions by Lopes et.al [5]. "

That sentences are moved to the discussion section

Limitation of work is missing

The limitations of our study are the assumption that the learning database is the frontal face and that the transform is constant. However, several other transformed images in the database have already been used for training.



The camera on the left is rotated around the Z axis, and the camera on the right is looking up at the actor from the bottom right. The above learning image damages the classification accuracy. This has already been explained in the Discussion and Conclusions section, but it needs to be highlighted. Therefore, we added two figures above (Figure 18) and added explanation with figure.

The discussion section fails to link the motivation of work. Expand your discussion section comparing your results with state-of-art results.

The most important motivation of our study was detail weighting which makes more efficient focus on each facial component. Our assumption is that an optimal grid map should place a high-depth grid around facial elements that are important for emotion changes (wrinkles in the corners of the eyes or mouth). A noteworthy was that the optimal grid map is not right and left symmetrical because it excludes the right and left symmetry of the actor's expression. This part is likely to require further experiments. We added this context in revised manuscript

The assumption of frontal face information being given with a few distortions are a limitation of this research. Because the images in the database used are not ideal fronts, the classification accuracy may have been diminished. In experimental results in section 4, an optimal grid map that quite close to assumption was derived as shown in below figure. However, this grid map is not right and left symmetrical because it excludes the right and left symmetry of the actor's expression. If this difference wasn't important, opposing girds would not been divided as our assumption. However, in the optimal grid map, it can be inferred that the difference in left and right information may affect classification accuracy. Several other transformed images in the database have already been used for training. Below figure shows two examples of that in ck+ database.

Better to separage discussion and conclusion.

We separated section 5. Discussion and Conclusion into 5. Discussion and 6. Conclusion with additional contents.