Context Enhancement Methodology for Action Recognition in Still Images

Jiarong He, Wei Wu(✉), and Yuxing Li

Inner Mongolia University, Hohhot 010021, Inner Mongolia, China
cswuwei@imu.edu.cn

Abstract. Action recognition in still images is a popular research topic in the field of computer vision, but it is to remain challenging due to the lack of motion information. Contextual information is a significant factor in the task of recognizing image action, which is inseparable from a predefined action class. And the existing research strategy does not ensure adequate use of contextual information. To address this issue, we propose a Contextual Enhancement Module (CEM) that combines the self-attention mechanism and the contextual attention mechanism. Specifically, the context enhancement module uses self-attention to learn pixel-level contextual information, after which separates the image into parts and uses contextual attention to learn region-level contextual information. In this way, the model can emphasize the significance of various pixels and regions in the image and significantly improve feature representation. We performed a lot of experiments on the PASCAL VOC 2012 Action dataset and the Stanford 40 Actions dataset. The results demonstrate that our method performs effectively, with the state-of-the-arts outcomes being obtained on both datasets.

Keywords: Action recognition · Attention mechanism · Contextual information

1 Introduction

Action recognition is a difficult study area in the world of computer vision and is widely applied in domains like as surveillance, robotics, human-computer interaction, and other areas [1]. The two categories of action recognition are image-based action recognition and video-based action recognition. However, recognizing actions in images is more challenging due to the lack of motion information, complex background, and high intra-class variance and low inter-class variance in some categories [2].

Images contain more information, such as human beings, interactive objects and scenes, which are composed of pixels. Humans can accurately distinguish these pieces of information, which indicates that there are certain connections between pixels of different information, we call these connections as the context information of images. Context information is one of the important clues in images, which is used in many image action recognition methods, however, most of the methods [7–10] consider from the perspective of multiple features fusion, and do not focus on the extraction of context

L. Iliadis et al. (Eds.): ICANN 2023, LNCS 14254, pp. 112–122, 2023.
https://doi.org/10.1007/978-3-031-44207-0_10

information. Only a few researchers have proposed the recognition method [11] using context information, but the experimental results are not satisfactory.

After achieving success in the field of natural language processing, attention mechanism [3] found widespread application in computer vision. Self-attention mechanism [4] is a special attention mechanism, which pays more attention to the key information contained in the input data itself. The self-attention mechanism assigns weights to each pixel in an image and then aggregates local features based on weighted summation. Therefore, we use the self-attentive mechanism to capture the correlation between pixels of an image to better describe global contextual information.

Regions can capture the object-parts relationships better, but they cannot be represented richly with only pixel-level contextual information. To truly describe an image, we must consider not only the spatial arrangement of the parts, but also their appearance and importance in distinguishing subtle differences. The context attention [5] can learn to emphasize potential representations of multiple regions, as well as encode spatial arrangements of various regions. It enables our model to selectively focus on more relevant integral regions to generate holistic context information.

Motivated by the observations above, in this paper, we propose a context enhancement module that uses a novel way to add two kinds of attention to the network, which can efficiently encode the spatial layout and visual appearance of parts. The contributions of this paper are summarized as follows:

- We propose a Context Enhancement Module (CEM). This module has a two-layer attention structure that combines a self-attention module and a contextual attention module to make the contextual information wealthy.
- We conduct experiments on the Stanford 40 Actions and PASCAL VOC 2012 Action datasets to demonstrate the effectiveness of CEM and the experiment parameters and network structure are introduced in detail. The results show that our methodology achieves the state-of- the-art performance.

2 Related Work

In 2006, Wang et al. [6] published the first paper on still image action recognition algorithms, and since then, with the rapid development of computer technology, especially the appearance of neural networks, more and more scholars have turned their attention to this aspect of deep learning.

In the field of deep learning, Gkioxari et al. proposed R*CNN [7], which incorporates contextual information as features in the recognition model. Zhao et al. [8] proposed a proposed method to arrange the features of different semantic parts in spatial order, arranging these features in a top-down order. Zhao et al. [9] proposed a method to improve human action recognition using semantic partial actions by merging local actions with contextual information.

With the occurrence of the attention mechanism, many authors began to try to bring it into their own models. Yan et al. [10] proposed a multi-branch attention network which has three branches, the scene attention branch, the target sub-region classification branch and the local region attention branch, thus capturing both global and local information. Zheng et al. [12] proposes a multi-stage deep learning method called Spatial Attention

based Action Mask Networks (SAAM-Nets). The model adds a spatial attention layer to the convolutional neural network to create a specific action mask for each image that has only an action label.

Additionally, some researchers are attempting to recognize actions in static images by using a variety of features. Ma et al. [13] proposed a new approach to action recognition by considering the relation between human and object as an important cue to enhance the features of action classification by computing the information of pair-wise relation between human and object. Wang et al. [14] proposed the pose enhanced relation module, which can extract the implicit relation between pose and human body output the pose enhanced relation feature with powerful representation capability. Surrounding objects information is also applied to strengthen the solution.

Most of the above methods only use convolutional neural networks to extract context information, and do not extract the context of images in depth. Some approaches conduct extensive research on context information, but the experimental results are not good. Compared with these methods, our proposed method can extract more detailed context information, which is conducive to improving the performance of recognizing actions in images.

3 Method

In this section, we introduce the model in detail. First, we'll go over the network's overall structure. Then, the two components of the Context Enhancement Module (CEM) are introduced in detail: the self-attention submodule and the context attention submodule.

3.1 Overview

Figure 1 shows the model's overall structure. First, ResNet-50 [15] is used for feature extraction, and the convolutional feature map of the last residual block in the network is retained. Then, the feature map is input into the Context Enhancement Module (CEM), where self-attention is employed to aggregates the contextual information of the overall image based on weighted summation, and contextual attention is used to enhance the feature representation of various regions and encode their spatial arrangement. Thus, the context enhancement module can emphasize the significance of individual pixels and regions in the image and obtain more detailed contextual information. Eventually, the dimension of the feature vectors is reduced by two fully connected layers to get the final recognition results. The next part gives a detailed presentation of the context enhancement module's structure.

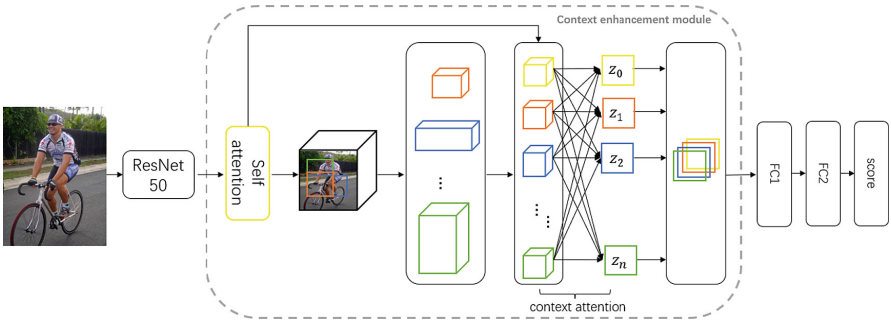

Fig. 1. Overview of our proposed methodology for action recognition in still images.

3.2 Context Enhancement Module (CEM)

The context enhancement module (CEM) structure is shown in Fig. 2. The module consists of two parts, namely self-attention submodule and contextual attention submodule. First, the image features are entered into the self-attention submodule, which learns pixel-level context information and generates a new feature map with self-attention weight. The context attention submodule takes this feature map as input and divides it into n integral regions, then extracts context information at the region level to produce n feature vectors. Finally, the module stacks these feature vectors to produce the final output feature map for the context enhancement module. As a result, the model could emphasize pixels and different-sized regions in the image as well learning contextual information in a hierarchical way.

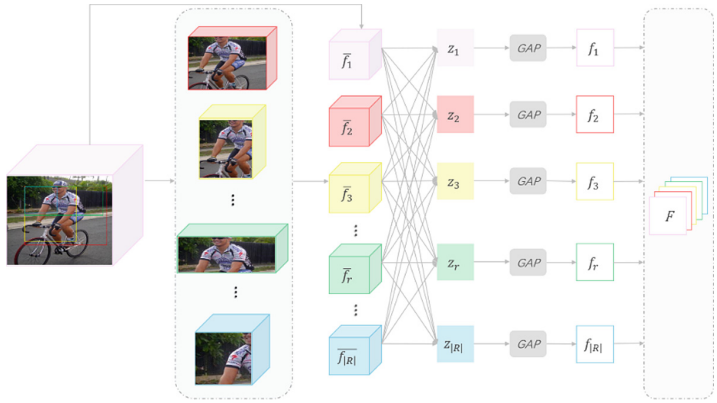

Fig. 2. The structure of the Context Attention Module (CEM)

Self-attention Submodule. In order to learn the relations among all pixels, we add a self-attentive module [16] to the model. $f(x)$, $g(x)$, $h(x)$ are 1×1 convolutions, and the output of $f(x)$ is transposed and multiplied with the output of $g(x)$. Through softmax,

we get an attention map θ_p, and multiply the attention map θ_p and $h(x)$ pixel by pixel to get the feature map o of self-attention. It is calculated as:

$$\theta_p = Softmax\left(\frac{g(x)f(x)^T}{\sqrt{d_k}}\right) \tag{1}$$

$$o = \theta_p * h(x) \tag{2}$$

where d_k denotes the number of feature dimensions. As a result, the model could not only learn global context information, but also focus on significant local information in the image.

Contextual Attention Submodule. For further extracting contextual information, we capture many regions with different roughness levels from the feature map, and the level of roughness is determined by the size of the region. The minimum region is $r(i, j, \Delta x, \Delta y)$, where Δx denotes the width and Δy denotes the height, located (top-left corner) in the i^{th} column and j^{th} row of the feature map o. We derive a set of regions by varying their widths and heights. The set of regions can be expressed as follows:

$$R = \{r(i, j, m\Delta x, n\Delta y)\} \tag{3}$$

Where $m, n = 1, 2, 3, \ldots$ and $i < i + m\Delta x \leq w,\ \mathrm{j} < \mathrm{j} + \Delta y \leq \mathrm{H}$. W and H denote the width and height of the feature map o, respectively. This method can obtain regions with different roughness in the feature map, so that the model can learn the subtle changes of different hierarchical structures in the image and obtain richer context information.

Since the size of region $r \in R$ is different, the goal is to represent these variable size regions $(X \times Y \times C) \rightarrow (w \times h \times C)$ with a fixed size feature vector, we process it using a bilinear pooling [17], usually a bilinear interpolation to achieve a differentiable image transformation. Let $T_\varphi(y)$ be the coordinate transformation of φ and $y = (i, j) \in R^2$ be the region coordinates with feature value $F(y) \in R^C$. The transformed image $\tilde{F}$ at the target coordinate $\tilde{y}$ is:

$$\tilde{F}(\tilde{y}) = \sum\nolimits_y F\big(T_\varphi(y)\big)K\big(\tilde{y}, T_\varphi(y)\big) \tag{4}$$

where $F\big(T_\varphi(y)\big)$ is the image indexing operation and is nondifferentiable; thus, the way gradients propagation through the network depends on the kernel $K(.,.)$. We use bilinear pooling to pool fixed size features $f_r(w \times h \times C)$ from all $r \in R$.

To obtain more detailed contextual information, fixed-size feature vectors are used as input to the contextual attention module [5] and contextual feature vector z_r as output. This module converts f_r to weighted versions of itself, conditional on the remaining feature mapping $f_{r\prime}(r, r\prime \in R)$. This allows our model to selectively focus on the more relevant integration regions to generate overall contextual information. It is calculated as:

$$g_{r,r'} = \tanh\Big(W_g(f_r) + W_{g'}\big(f_{r'}\big) + b_g\Big) \tag{5}$$

$$\alpha_{r,r'} = softmax\Big(W_\alpha g_{r,r'} + b_\alpha\Big) \tag{6}$$

$$z_r = \sum_{r'=1}^{R} \alpha_{r,r'} f_{r'} \tag{7}$$

where W_g, $W_{g'}$ are the weight matrices of f_r and $f_{r\prime}$, W_α is the weight matrix of their nonlinear combination, and b_α and b_g are the bias matrices. The attention element $\alpha_{r,r\prime}$ captures the similarity between the feature maps f and $f_{r\prime}$ of regions r and $r\prime$. The attention focused context vector z_r determines the strength of f_r in focus conditioned on itself and its neighborhood context. This applies to all integral regions r.

In order to improve the extensibility of the model and reduce the computational complexity of the model, we use global average pooling to integrate the spatial information of the feature vector $z_r(r = 1, 2, 3, \cdots) \in \mathbb{R}^{w \times h \times C}$ and obtain the context feature $f_r \in \mathbb{R}^{1 \times C}$. To create the context attention sub-module's final output vector $F \in \mathbb{R}^{|\mathcal{R}| \times C}$, all of feature vectors f_r are finally stacked.

4 Experiments

In this section, we first provide a description of the experimental datasets and parameter settings, then compare our experimental results with the state-of-the-art models, and finally perform ablation experiments to prove the effectiveness of our proposed model.

4.1 Datasets and Evaluation Metric

We use the PASCAL VOC 2012 Action [18] dataset and the Stanford 40 Actions [19] dataset to train and evaluate the image action recognition task.

The PASCAL VOC 2012 Action dataset, which contains 9157 images covering 10 categories of actions. For training and validation, 400–500 images from each category in the dataset are used, and the remaining images are used for testing. The Stanford 40 Actions dataset consists of 9532 images total, separated into 40 classes of actions, with 100 pictures every class used for training and the rest images used for testing. The two datasets are split similarly to other methods that are currently in use in the field, allowing for a performance comparison with those.

For action recognition in images, we measure the performance by Average Precision (AP) and mean Average Precision (mAP). Average Precision (AP) is used to measure the performance of the model on each category, and mean Average Precision (mAP) is used to measure the overall performance of the model.

4.2 Experimental Setup

In our experiments, we set the input image size to 224 × 224 and the training epoch to 100 on all datasets. we utilize stochastic gradient descent (SGD) [20] with a momentum of 0.9 and a learning rate of 0.0001 to optimize the model during the training period. The entire model is constructed using the Tensorflow framework and trained on single NVIDIA Tesla P40 GPU.

4.3 Comparisons with the State-of-the-Art Models

In this section, we show the result of other state-of-the-art methods to provide a comprehensive perspective on the performance of our proposed model.

We firstly evaluate our model on the Pascal VOC 2012 Action dataset. The results and comparison with state-of-the-art approaches on the validation and test set are respectively shown in Tables 1 and 2. On the validation and test sets, the mAP of our method achieves 93.1% and 94.2%, which is the State-of-the-art result among all methods. Especially on the test set, our approach significantly improved the AP values for the categories of 'playing instrument', 'using computer' and 'walking' etc.

Table 1. Performance comparison on the PASCAL VOC 2012 Action validation set

Method	Jumping	Phoning	Playing instrument	Reading	Riding bike	Riding horse	Running	Taking photo	Using computer	Walking	Mean AP
R*CNN [7]	87.7	80.1	94.8	81.1	**95.5**	97.2	87.0	84.7	94.6	70.1	87.3
Yan et al. [10]	87.8	78.4	93.7	81.1	95.0	97.1	86.0	85.5	93.1	73.4	87.1
Ma et al. [13]	89.2	**89.8**	96.5	87.6	98.2	99.1	**92.3**	**91.6**	95.2	79.2	91.9
Zhao et al. [9]	89.6	86.9	94.4	88.5	94.9	97.9	91.3	87.5	92.4	76.4	90.0
Ours	**92.4**	84.5	**98.8**	**92.7**	**95.5**	**99.8**	91.6	91.2	**98.4**	**85.5**	**93.1**

Table 2. Performance comparison on the PASCAL VOC 2012 Action test set

Method	Jumping	Phoning	Playing instrument	Reading	Riding bike	Riding horse	Running	Taking photo	Using computer	Walking	Mean AP
R*CNN [7]	91.5	84.4	93.6	83.2	96.9	98.4	93.8	85.9	92.6	81.8	90.2
Yan et al. [10]	92.7	86.0	93.2	83.7	96.6	98.8	93.5	85.3	91.8	80.1	90.2
Ma et al. [13]	91.1	89.8	95.4	87.7	98.6	98.8	**95.4**	**91.4**	95.8	84.3	92.8
Zhao et al. [9]	95.0	**92.4**	97.0	88.3	**98.9**	99.0	94.5	91.3	95.1	87.0	93.9
Ours	**96.6**	89.5	**99.1**	**91.9**	97.8	**99.2**	91.4	87.7	**98.6**	**90.6**	**94.2**

We further evaluate the proposed model on the Stanford 40 Actions dataset. As shown in Table 3, The mAP of our proposed method is 95.0%, achieving the state-of-the-art performance. In particularly, The approach [13, 14, 21] focuses more on recognizing the features of the interaction relationship between people and objects, but our approach

evaluates from the aspect of context information, increasing the mAP value by 1.8–5.5%. The method [7, 10, 23] approaches the problem from a similar perspective as ours, but these methods perform worse than ours, with a performance difference of 1.2–4.3%.

Table 3. Performance comparison on the Stanford 40 Actions validation set

Method	Networks	Mean AP(%)
Mi et al. [21]	ResNet-101	89.5
Yan et al. [10]	VGG-16	90.7
R*CNN [7]	VGG-16	90.9
Zhao et al. [9]	ResNet-50	91.2
Ma et al. [13]	ResNet-50	93.1
Wang et al. [14]	ResNet-50	93.2
Wu et al. [22]	ResNet-50	93.7
Li et al. [23]	ResNet-50	93.8
Ours	ResNet-50	**95.0**

4.4 Ablation Study and Analysis

In this section, we conducted detailed ablation experiments on two datasets to demonstrate the effectiveness of our proposed method.

Table 4. Ablation study on the two datasets

Method	ResNet-50	Context Enhancement Module		Mean AP(%)	
		Self-Attention Submodule	Context Attention Submodule	PASCAL VOC 2012	Stanford 40
1	✓			70.3	78.8
2	✓	✓		73.7	80.5
3	✓		✓	93.9	94.6
4	✓	✓		**94.2**	**95.0**

Firstly, we explored the impact of the model's three components on the experimental results, and the data are shown in Table 4. As shown in the table, the contextual attention submodule plays a much significant role than the self-attention submodule. The experimental results of adding the Context Enhancement Module into the model is the greatest, with mAP of 94.2% and 95.0%, respectively, confirming that the proposed hierarchical learning approach of pixel-level and region-level context information is beneficial for recognizing action in still image.

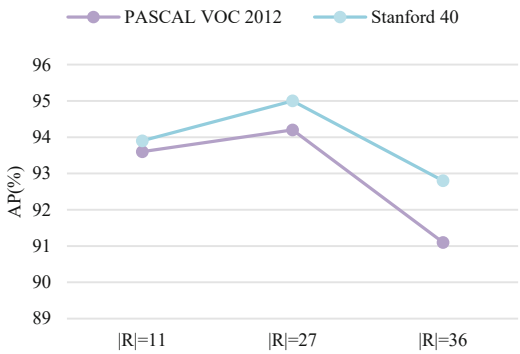

Fig. 3. Experimental results of different number $|R|$ of integral regions

Figure 3 illustrates the effect of the number $|R|$ of Integral Regions on model performance. There are 10, 26, and 35 integral regions that can be obtained by altering the values of m and n in Formula (3) of Sect. 3, including the input image, the total number of regions is 11, 27, and 36. When the number of regions was increased to 27, the PASCAL VOC 2012 and Stanford 40 Actions datasets had the highest mAP values, proving that the region provides the most contextual information in this setting. When the number of regions is 36, the mAP values of the PASCAL VOC 2012 and Stanford 40 actions datasets are the lowest, which means that the different regions overlapped more at this time and the information contained in the feature maps was in an oversaturated state, causing the performance of the model to decrease. This experiment shows that when there are 27 regions, the model performs best on both datasets.

5 Conclusions

This paper presents a novel action recognition model based on contextual information. Context information is an important clue of image activity recognition, but the existing methods do not make full use of it, resulting in poor recognition effect of static images. We created a multiple-attention fusion strategy to solve this problem, which build the context-enhanced modules by applying attention mechanisms in order to gather more valuable contextual information for enhancing feature representation. Experimental results demonstrates that our method performs better than the state-of-the-art models on PASCAL VOC 2012 and Stanford 40 Actions datasets.

Acknowledgement. This work is supported by the Inner Mongolia Science and Technology Project (No. 2021GG0166).

References

1. Zhu, Y., et al.: A comprehensive study of deep video action recognition. arXiv preprint arXiv: 2012.06567 (2020)
2. Girish, D., Singh, V., Ralescu, A.: Understanding action recognition in still images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 370–371 (2020)
3. Dosovitskiy, A., et al.: An image is worth 16 x 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
5. Behera, A., Wharton, Z., Hewage, P.R., Bera, A.: Context-aware attentional pooling (cap) for fine-grained visual classification. Proc.AAAI Conf. Artif. Intell. **35**(2), 929–937 (2021)
6. Wang, Y, et al.: Unsupervised discovery of action classes. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 2. IEEE (2006)
7. Gkioxari, G., Girshick, R., Malik, J.: Contextual action recognition with r* CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1080–1088 (2015)
8. Zhao, Z., Ma, H., Chen, X.: Semantic parts based top-down pyramid for action recognition. Patt. Recogn. Lett. **84**, 134–141 (2016)
9. Zhao, Z., Ma, H., You, S.: Single image action recognition using semantic body part actions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3391–3399 (2017)
10. Yan, S., Smith, J.S., Lu, W., Zhang, B.: Multibranch attention networks for action recognition in still images. IEEE Trans. Cognitive Dev. Syst. **10**(4), 1116–1125 (2017)
11. Zhu, H., Hu, J.F., Zheng, W.S.: Learning hierarchical context for action recognition in still images. In: Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, 21–22 September, 2018, Proceedings, Part III 19, pp. 67–77 (2018)
12. Zheng, Y., Zheng, X., Lu, X., Wu, S.: Spatial attention based visual semantic learning for action recognition in still images. Neurocomputing **413**, 383–396 (2020)
13. Ma, W., Liang, S.: Human-object relation network for action recognition in still images. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2020)
14. Wang, J., Liang, S.: Pose-enhanced relation feature for action recognition in still images. In: Þór Jónsson, B., et al. MultiMedia Modeling. MMM 2022. Lecture Notes in Computer Science, vol. 13141. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98358-1_13
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363 (2019)
17. Yu, C., Zhao, X., Zheng, Q., Zhang, P., You, X.: Hierarchical bilinear pooling for fine-grained visual recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 574–589 (2018)
18. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal Visual Object Classes (voc) challenge. Int. J. Comput. Vision **88**, 303–338 (2010)
19. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: 2011 International Conference on Computer Vision, pp. 1331–1338 (2011)
20. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv: 1609.04747(2016)

21. Mi, S., Zhang, Y.: Pose-guided action recognition in static images using lie-group. Appl. Intell. 1–9(2022)
22. Wu, W., Yu, J.: An improved deep relation network for action recognition in still images. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2450–2454 (2021)
23. Li, Y., Li, K., Wang, X.: Recognizing actions in images by fusing multiple body structure cues. Patt. Recogn. **104**, 107341 (2020)