TAPoseNet: Teeth Alignment based on Pose estimation via multi-scale Graph Convolutional Network*

Qingxin Deng[1], Xunyu Yang[1], Minghan Huang[1], Landu Jiang[2], and Dian Zhang[1]

[1] College of Computer Science and Software Engineering, Shenzhen University, Nanhai Avenue 3688 Shenzhen, China zhangd@szu.edu.cn
[2] The Hongkong University of Science and Technology (Guangzhou), Duxue Road 1, Guangzhou, China landujiang@hkust-gz.edu.cn

Abstract. Teeth alignment plays an important role in orthodontic treatment. Automating the prediction of teeth alignment target can significantly aid both doctors and patients. Traditional methods often utilize rule-based approach or deep learning method to generate teeth alignment target. However, they usually require extra manual design by doctors, or produce deformed teeth shapes, even fail to address severe misalignment cases. To tackle the problem, we introduce a pose prediction model which can better describe the space representation of the tooth. We also consider geometric information to fully extracted features of teeth. In the meanwhile, we build a multi-scale Graph Convolutional Network(GCN) to characterize the teeth relationships from different levels (global, local, intersection). Finally the target pose of each tooth can be predicted and so the teeth movement from the initial pose to the target pose can be obtained without deforming teeth shapes. Our method has been validated in clinical orthodontic treatment cases and shows promising results both qualitatively and quantitatively.

Keywords: Deep learning · 3D point cloud · Orthodontic treatment.

1 Introduction

Teeth alignment is a critical concern in dentistry, satisfying the human requirement to become more beautiful and healthy [2]. As a result, the demand for orthodontic treatment is rising dramatically. In orthodontic treatment, the determination of the teeth alignment target is crucial, as it directly determines the subsequent treatment plans and the design of the orthodontic appliance. While computer-aided modeling techniques such as intra oral scans have revolutionized the field of orthodontics and provided increased patient comfort, they do require substantial time commitment from doctors and orthodontists to determine teeth alignment target. Therefore, it is essential to develop a fully automated system

* The corresponding author is Dian Zhang.

to determine an optimal teeth arrangement target. A system like this would not only alleviate the burden of manual operations for dentists but would also enhance the communication effectiveness with patients, by allowing patients to envisage the results of their future dental arrangements.

Existing methods to address this challenge can be roughly divided into three categories: 1) Automatic teeth alignment based on expertise rules. These methods usually require extensive manual intervention from doctors, and typically require prior information such as teeth landmarks. Cheng et al. [5] propose an accurate teeth arrangement system with complete teeth model. But, this system requires manual intervention from doctors to set the alignment target of incisors. Deng et al. [7] propose an automatic approach for maxilla and mandible alignment. However this alignment requires many pre-defined teeth landmarks. 2) Automatic teeth alignment based on generative model. These methods are mainly used for 2D images, with some designed for 3D models that don't need any pre-segmentation of teeth. However, this approach might lead to severe distortions in the 3D models of teeth. Chen et al. [4] present a method to predict the visual outcome of orthodontic treatment in a portrait image via latent style code manipulation. Yang et al. [11] propose a system which takes a frontal face image of a patient along with a corresponding 3D teeth model as input and generates a facial image with aligned teeth. Zhang et al. [17] present the first parametric 3D morphable dental model for both teeth and gum which can be used to smoothly interpolate between pre-orthodontic teeth and post-orthodontic teeth. However, the size and shape of the teeth might be changed during interpolation. 3) Automatic teeth alignment based on regressing the transformaion matrices for each tooth. PSTN [10] proposed by Li et al. inspired by [9] uses PointNet [12] and PointNet++ [13] for global and local features extraction and then directly regresses the transformation matrices. TANet [16] proposed by Wei et al. uses graph-based feature propagation module to update features extracted by PointNet [12] to solve the 6-DOF pose prediction problem of each tooth. Wang et al. [14] uses anatomical landmark constraints to improve tooth alignment results instead of directly regressing tooth motion. However, these methods require pairs of pre-orthodontic and post-orthodontic teeth models to train the model and it is difficult to handle severe misalignment cases.

In order to address aforementioned challenges, we propose a system named TAPoseNet, which can fully automate the prediction of post-orthodontic teeth alignment target without any deformation of teeth. TAPoseNet includes a Teeth Pose Estimation module based on DGCNN [15] to explicitly estimate the pose of each tooth in dental arch, paired with a Geometric Information Extraction Module that extracts each tooth's geometric information. Afterwards the extracted features are fed into Teeth alignment target prediction module to predict the post-orthodontic pose of each tooth. Finally, transformation matrices from pre-orthodontic poses to post-orthodontic poses are used for transitioning the pre-orthodontic teeth to the predicted post-orthodontic arrangement.

The contributions of our work are as follows: 1) We present a method to automatically predict the post-orthodontic teeth alignment target based on initial

teeth pose estimation and target teeth pose prediction, without deforming teeth shapes. 2) To the best of our knowledge, we introduce the first deep learning based method to estimate the pose of teeth, which can better describe the space representation of the tooth and contribute to the prediction of post-orthodontic teeth arrangement target with clinical interpretability. 3) We build a multi-scale Graph Convolutional Network (GCN) to characterize the spatial relationships of teeth in multi-scale from different levels (global, local, intersection).

2 Method

TAPoseNet is composed of two major components: 1) Teeth features extraction module and 2) Teeth alignment target prediction module. (Fig. 1)

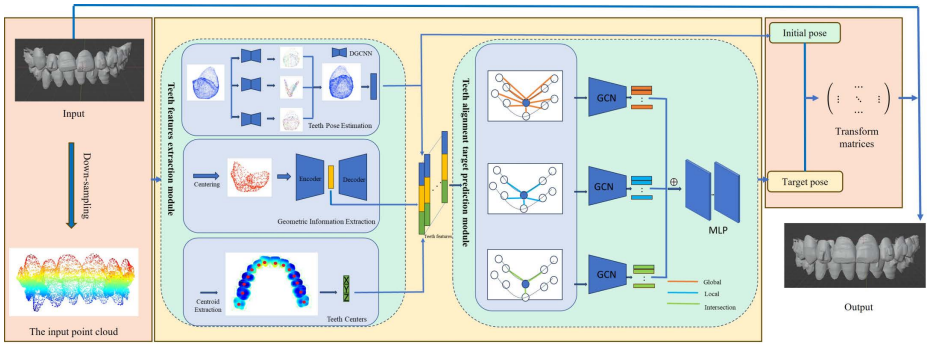

Fig. 1. The overall network architecture of our method.

The central idea is to approach the prediction of teeth alignment target as a problem of predicting the target pose of each tooth given the initial pose of each tooth. To accurately estimate the initial pose of teeth, we propose a teeth pose estimation module to regress the local coordinates of teeth. For the geometric information of teeth, we model it as a latent code that can be effectively extracted from the input teeth point cloud. Finally, 3 GCNs with different adjacency matrices are proposed to aggregating the teeth features in multi-scale.

TAPoseNet operates as follows. The input of our network is a segmented and classified teeth crown point cloud $P = \left\{P_v \subseteq R^{N_t \times 3} \mid v \in \mathcal{V}\right\}$ down-sampled from the pre-orthodontic teeth model $T = \{T_v \mid v \in \mathcal{V}\}$, where $\mathcal{V}$ denotes the set of tooth labels which are assigned according to FDI two digit notation for permanent teeth and N_t is the number of sampled points of each tooth. The teeth features extraction module mainly extracts the geometric information *geocode* and the pre-orthodontic pose $Pose_{pre}$ of each tooth. The post-orthodontic teeth alignment target prediction module predicts the post-orthodontic pose of each tooth $Pose_{post}$ by aggregating the extracted tooth features in multi-scale. Applying the transformation matrices from initial poses to predicted poses to the

pre-orthodontic teeth model, we can obtain the post-orthodontic teeth model $T' = \{T'_v \mid v \in \mathcal{V}\}$.

2.1 Teeth features extraction module

The features of a tooth consists of two parts: the pose of the tooth and the geometric information of the tooth. The pose of a tooth is a representation of its position and posture in the local coordinate system L relative to the world coordinate system W, described by rotation and translation in three-dimensional space. Assuming that there is a point p on a specific tooth, the coordinate of p in L is p_L, the coordinate of p in W is p_W, the rotation matrix $R \in SO(3)$ is for the rotation and the vector C is for the translation, the transformation from p_W to p_L can be expressed as:

$$p_L = R \cdot (p_W - C) \tag{1}$$

Consequently, the pose of the tooth is represented by the quaternion rotation R^{-1} which is a four-dimensional vector and the centroid of the tooth C which is a three-dimensional vector.

To obtain the local coordinate system of a specific tooth, we propose a 3 head architectures of DGCNN [15] for predicting the x, y, z coordinate of the tooth in the local coordinate system. (Fig.1) For the stability of training, we discretize the local coordinates to transform the coordinate regression problem into a point cloud classification problem. Specifically, the input of this module is a point cloud of a segmented tooth $P_v \subseteq R^{N_t \times 3}$, the output of a head is of size $N_t \times Num_{class}$, where Num_{class} is the number of categories in the x (y, z) direction.

In the process of determining the target position of orthodontic treatment, the abstract pose of teeth cannot be solely considered. The shape of different teeth, surface texture, bite groove and other information have a significant impact on the teeth alignment target determination. In order to effectively extract the geometric shape information of teeth surfaces, we trained a deep Autoencoder (AE) to encode the geometric shape features of teeth [1]. Specifically, for the input tooth point cloud P_v, the encoder based on PointNet [12] outputs a latent feature $geocode \subseteq R^{N_{latent}}$. The decoder based on MLP then outputs the reconstructed point cloud P'_v given the latent feature as input. Using Chamfer distance (2) as the loss function for training,

$$Loss\left(P_v, P'_v\right) = \sum_{x \in P_v} \min_{y \in P'_v} \|x - y\|_2^2 + \sum_{y \in P'_v} \min_{x \in P_v} \|x - y\|_2^2 \tag{2}$$

the encoder can effectively extract the geometric shape information of the tooth. At the inference stage, we only retain the encoder part for extracting the geometric information of the tooth point cloud.

2.2 Post-orthodontic teeth alignment target prediction module

We predict the post-orthodontic pose $Pose_{post}$ of each tooth using a deep model based on the pre-orthodontic pose $Pose_{pre}$ and geometric information *geocode* extracted from the teeth features extraction module. Specifically, we concatenate the initial pose vector containing a 4 dimensional quaternion and a 3 dimensional centroid coordinate with geometric feature vector which is a 100 dimensional latent vector as input $X = (Pose_{pre}, geocode)$ for each tooth. Therefore the input embedding is in shape of $N \times (7+100)$, N being the total number of teeth of a patient (usually 28). This embedding is then fed into Teeth alignment target prediction module composed of GCN-based encoder and MLP-based decoder. The encoder employs 3 GCNs with different adjacency matrices constructed from different spatial dependencies between teeth in the dental arch G_{global}, G_{local} and $G_{intersection}$. Specifically, by examining each tooth as a node within the dental arch, three types of adjacency matrices are formulated: 1) Global adjacency matrix. This matrix treats the dental arch as a fully interconnected graph so that every tooth connects with all others. This allows the network to extract and interpret the holistic arch shape information. 2) Local adjacency matrix. In this matrix, each tooth is linked not only with the opposing tooth but also with the adjacent ones, as well as their opposite adjacent teeth. This linkage allows the network to discern the localized crowded occlusion within the dental arch. 3) Intersection adjacency matrix, every tooth is connected to both the opposing tooth and the adjacent ones. This structure primarily serves to prevent potential collisions between teeth. The outputs of the GCNs are concatenated and then fed into the MLP-based decoder D. The output of the decoder is in shape of $N \times 7$, which indicates the predicted post-orthodontic pose of each tooth of a patient.

$$Pose_{post} = D(G_{global}(X) \oplus G_{local}(X) \oplus G_{intersection}(X)) \tag{3}$$

To train the network, we adopt a loss function to compute the difference between $Pose_{post}$ and the ground truth pose $Pose_{gt}$.

$$Loss = Loss_{rotation} + Loss_{translation} \tag{4}$$

Specifically, $Loss_{rotation}$ measures the cosine similarity between prediction quaternion posture and ground truth quaternion posture, $Loss_{translation}$ calculates the distance between prediction position of tooth and ground truth position of tooth. Given the poses before and after orthodontic treatment, we can calculate the transformation matrices $(Trans_{pose_{pre} \to pose_{post}})_v$ for each tooth. Therefore the post-orthodontic teeth model can be obtained by applying the transformation matrices to each tooth of the patient.

$$T' = \left\{ (Trans_{pose_{pre} \to pose_{post}})_v \times T_v | v \in \mathcal{V}, T_v \subseteq T \right\} \tag{5}$$

3 Experiments and Results

3.1 Dataset

Our experiment was conducted on a dataset of clinical orthodontic cases sourced from a dental hospital. This dataset comprises post-orthodontic oral scan data from 50 patients, along with pre- and post-orthodontic treatment oral scan data from 25 pairs. Given the potential discrepancies in the orientations of oral scan data due to the varying features of oral scanning devices, we employed the ICP registration method [6] to standardize the orientation across all dataset. Then we use the harmonic field method [18, 3] to segment each tooth from the oral scan and remove the gingival part, leaving only the crown part. Finally, we label each tooth crown using the FDI digit. For network training, we use post orthodontic oral scan data from 50 patients. During training, we utilized data augmentation methods (e.g., randomly rotate or translate) in each epoch to reverse-generate diverse initial poses (pre-orthodontic) as input for each case in the training set with ideal target poses (post-orthodontic), reflecting different orthodontic symptoms. For validation and testing we utilize 25 pairs of oral scan data that were gathered before and after orthodontic treatment (10 for validation and 15 for testing). In the inference stage, initial pose of each tooth from the input is estimated by the pre-trained Teeth Pose Estimation module, so that we can predict the target pose.

3.2 Implementation and evaluation methods

The implementation detail of TAPoseNet is described below. We randomly down sample 1024 points from each tooth crown of a patient. During the teeth pose estimation, The output dimension of each DGCNN network is 32. In teeth alignment target prediction module, the MLP-based decoder includes several shared FC layers, a squeeze-and-excitation(SE) block [8] with reduction ratio 4, and skip connection.

We trained the model for 4000 epochs with a batch size of 2 using Adam opimizer. The learning rate was initialized as 1e-4 and use the cosine learning rate scheduler with the minimun learning rate set to be 1e-6. The models were trained on an NVIDIA RTX-2080 Ti.

The prediction accuracy of TAPoseNet was evaluated quantitatively and qualitatively by comparing it to representative methods PSTN [10], TAligNet [11] and TANet [16]. Given that in clinical practice, the pre- and post-orthodontic oral scan teeth models typically do not share the same coordinate system and often display inconsistencies in the model's vertex counts, we adopted the Chamfer Distance(CD) as one of our evaluation metrics. This metric evaluates the distance between the predicted post-orthodontic teeth model and the ground truth post-orthodontic teeth model post-rigid registration implemented through the Iterative Closest Point (ICP) algorithm [6]. Additionally, the measurement of matrix similarity can act as an accuracy index since the 3D transformation of teeth are described by spatial transformation matrix. Therefore, we randomly

disarrange the post-orthodontic teeth and calculate the commonly used cosine similarity accuracy (CSA) to measure the difference between the generated transformation matrix and its ground truth.

3.3 Results

The results of quantitative evaluation of TAPoseNet mainly focus on prediction accuracy of teeth alignment target. We compared our methd with representation methods, TANet [16], PSTN [10] and TAligNet [11] for comparison. Significantly, TANet [16] and PSTN [10] are fully automatic method without any prior information, while TAligNet [11] and the Post-orthodontic teeth alignment target prediction module of our TAPoseNet need the input of the initial pose information of the teeth. To ensure the fairness, we use the teeth pose estimation module of our TAPoseNet to estimate the pose of teeth automatically. The results are shown in Table 1. As shown in the table, our network achieves the lowest Chamfer Distance and the highest Cosine Similarity Accuracy.

Table 1. Result comparison of different methods.

Methods		Chamfer Distance (mm) ↓	CSA% ↑
non-pose	PSTN	0.68667	84.52
	TANet	0.68533	84.54
pose-based	TAligNet	0.62281	85.94
	TAPoseNet	**0.60457**	**86.77**

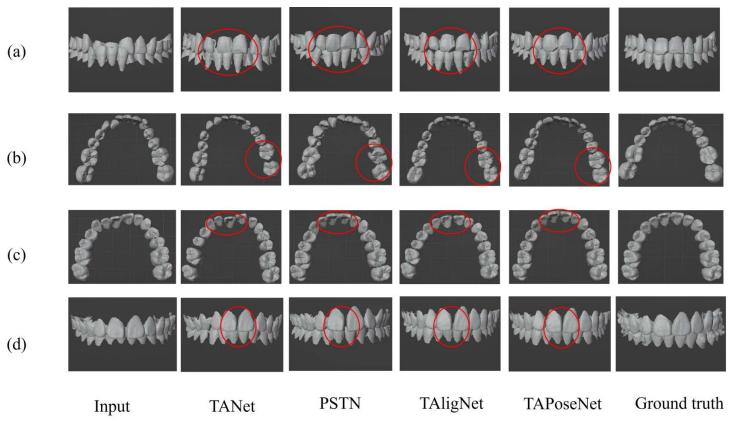

Fig. 2. Visualization result of comparison experiments with other methods. (a) is the front view of the whole dental model of the first patient, (b) is mandible of the first patient (c) is maxilla of the first patient. (d) is the front view of the second patient.

Fig. 2 presents the results of different methods tested on pre-orthodontic teeth models. Taking the patient with severe misaligned teeth as an example(the first three rows). From the front view(a), we can see that TANet and PSTN which directly regress the motion of each tooth cannot tackle the severe misalignment. The positions and postures of some teeth are obviously still misaligned. TAlignNet based on pose prediction without multi-scale feature aggregation performs better in this case, but the relationship between maxilla and mandible is unreasonable. From the mandible(b) and maxilla(c), we can see that TAPoseNet provides the most optimal results, particularly when managing the relationships between teeth. Taking the patient with mild underbite problem as an example(the last row), we focus on : 1) The distance on the vertical direction where the upper row of teeth covers the lower row of teeth; 2) Whether the midline of the teeth is aligned or not. We can see that TAPoseNet performs best in these two aspects.

From the quantitative evaluation and qualitative evaluation, we can learn that the methods using pose information perform better than methods that directly regress the motion of teeth. Therefore, the accuracy of pose estimation is essential. We also conduct quantitative evaluation (Table 2) using the mean point-wise distance and maximum point-wise distance as metric on our teeth pose estimation module. We use the Oriented Bounding Box (OBB) and the Axis Aligned Bounding Box (AABB) as visualization for qualitative evaluation (Fig. 3(a)). In practice, the error of the estimated pose is acceptable(Fig. 3(b));

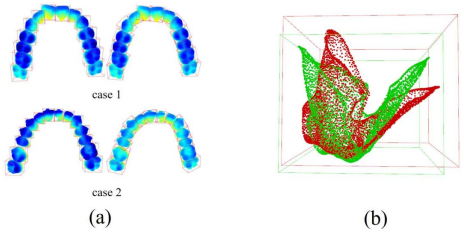

Fig. 3. The visualization of teeth pose estimation. (a) is the axis aligned bounding box and oriented bounding box based on estimated pose of the dental model. (b) is a comparison between the predicted teeth pose (green) and the ground truth teeth pose (red) whose mean point-wise distance is 1.52326

Table 2. The mean point-wise distance and maximum point-wise distance between the predicted teeth pose and ground truth teeth pose.

teeth categories	Mean point-wise distance (mm)	Maximum point-wise distance (mm)
Incisor	0.39352	0.52463
Canine	1.00160	1.52326
Premolar	1.14999	1.43654
Molar	0.52546	0.82379

4 Discussions and Conclusions

We proposed a deep learning-based framework, TAPoseNet to predict teeth alignment target. The quantitative evaluation and qualitative evaluation demonstrate the effectiveness of TAPoseNet in orthodontic treatment planning. An integral component of TAPoseNet is the teeth pose estimation module, which automatically estimates teeth poses, significantly contributing to various facets of orthodontic treatment planning. For future work, we need to consider the occlusion of the upper and lower jaw. Additionally, missing teeth or wisdom teeth cannot be handled in our work, which will be handled in future work.

Acknowledgments. This work was supported in part by Stable Support Project of Shenzhen (Project No. 20231122145548001), the JCYJ under Grant 20220531091407016 and the HKUST(GZ)-ROP2023056.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: International conference on machine learning. pp. 40–49. PMLR (2018)
2. Andrews, L.F.: The six keys to normal occlusion. Am J orthod **62**(3), 296–309 (1972)
3. Au, O.K.C., Zheng, Y., Chen, M., Xu, P., Tai, C.L.: Mesh segmentation with concavity-aware fields. IEEE Transactions on Visualization and Computer Graphics **18**(7), 1125–1134 (2011)
4. Chen, B., Fu, H., Zhou, K., Zheng, Y.: Orthoaligner: image-based teeth alignment prediction via latent style manipulation. IEEE Transactions on Visualization and Computer Graphics (2022)
5. Cheng, C., Cheng, X., Dai, N., Liu, Y., Fan, Q., Hou, Y., Jiang, X.: Personalized orthodontic accurate tooth arrangement system with complete teeth model. Journal of medical systems **39**, 1–12 (2015)
6. Chetverikov, D., Svirko, D., Stepanov, D., Krsek, P.: The trimmed iterative closest point algorithm. In: 2002 International Conference on Pattern Recognition. vol. 3, pp. 545–548. IEEE (2002)
7. Deng, H., Yuan, P., Wong, S., Gateno, J., Garrett, F.A., Ellis, R.K., English, J.D., Jacob, H.B., Kim, D., Xia, J.J.: An automatic approach to reestablish final dental occlusion for 1-piece maxillary orthognathic surgery. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 345–353. Springer (2019)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
9. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)

10. Li, X., Bi, L., Kim, J., Li, T., Li, P., Tian, Y., Sheng, B., Feng, D.: Malocclusion treatment planning via pointnet based spatial transformation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 105–114. Springer (2020)
11. Lingchen, Y., Zefeng, S., Yiqian, W., Xiang, L., Kun, Z., Hongbo, F., Zheng, Y.: iorthopredictor: model-guided deep prediction of teeth alignment. ACM Transactions on Graphics **39**(6), 216 (2020)
12. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
13. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
14. Wang, C., Wei, G., Wei, G., Wang, W., Zhou, Y.: Tooth alignment network based on landmark constraints and hierarchical graph structure. IEEE Transactions on Visualization and Computer Graphics (2022)
15. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019)
16. Wei, G., Cui, Z., Liu, Y., Chen, N., Chen, R., Li, G., Wang, W.: Tanet: towards fully automatic tooth arrangement. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 481–497. Springer (2020)
17. Zhang, C., Elgharib, M., Fox, G., Gu, M., Theobalt, C., Wang, W.: An implicit parametric morphable dental model. ACM Transactions on Graphics (TOG) **41**(6), 1–13 (2022)
18. Zou, B.j., Liu, S.j., Liao, S.h., Ding, X., Liang, Y.: Interactive tooth partition of dental mesh base on tooth-target harmonic field. Computers in biology and medicine **56**, 132–144 (2015)