d/Collection

치의학박사학위논문

Comparison of accuracy and computational performance between the latest machine learning algorithms for automated cephalometric landmark identification – YOLOv3 vs SSD

두부계측방사선 사진 계측점 자동 식별의 최신 기계 학습 알고리즘 간 정확도 및 연산 성능 비교 연구 – YOLOv3 vs SSD

2019년 8월

서울대학교 대학원

치의과학과 치과교정학 전공

박 지 훈

치의학박사학위논문

Comparison of accuracy and computational performance between the latest machine learning algorithms for automated cephalometric landmark identification – YOLOv3 vs SSD

지도교수 이 신 재

이 논문을 치의학박사 학위논문으로 제출함
2019 년 6 월

서울대학교 대학원
치의과학과 치과교정학 전공
박 지 훈

박지훈의 치의학박사 학위논문으로 인준함

2019 년 7 월

위 원 장   (인)

부위원장   (인)

위   원   (인)

위   원   (인)

위   원   (인)

- ABSTRACT -

Comparison of accuracy and computational performance between the latest machine learning algorithms for automated cephalometric landmark identification – YOLOv3 vs SSD

Ji-Hoon Park, BS, DDS

Department of Orthodontics, Graduate School, Seoul National University

*(Directed by Professor **Shin-Jae Lee**, DDS, MSD, PhD, PhD)*

Introduction: The purpose of this study was to compare two of the latest deep learning algorithms for automatic identification of cephalometric landmarks in their accuracy and computational efficiency. This study uses two different algorithms for automated cephalometric landmark identification with an extended number of landmarks: 1) You-Only-Look-Once version 3 (YOLOv3) based method with modification, and 2) the Single Shot Detector (SSD) based method.

Materials and methods: A total of 1,028 cephalometric radiographic images were selected as learning data that trained YOLOv3 and SSD methods. The number of target labelling was 80 landmarks. After the deep learning process, the algorithms were tested using a new test data set comprised of 283 images. The accuracy was determined by measuring the

mean point-to-point error, success detection rate (SDR), and visualized by drawing 2-dimensional scattergrams. Computational time of both algorithms were also recorded.

Results: YOLOv3 algorithm outperformed SSD in accuracy for 38/80 landmarks. The other 42/80 landmarks did not show a statistically significant difference between YOLOv3 and SSD. Error plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency. Mean computational time spent per image was 0.05 seconds and 2.89 seconds for YOLOv3 and SSD, respectively. YOLOv3 showed approximately 5% higher accuracy compared with the top benchmarks in the literature.

Conclusions: Between the two algorithms applied, YOLOv3 seems to be promising as a fully automated cephalometric landmark identification system for use in clinical practice.

CONTENTS

Abstract

Contents

I. INTRODUCTION

The use of machine learning techniques in the field of medical imaging is rapidly evolving.[1-4] Attempts to apply machine learning algorithms in orthodontics are also increasing. Some of the major applications currently utilized are automated diagnostics,[5] data mining,[6,7] and landmark detection.[8-10] Inconsistency in landmark identification has been known to be a major source of error in cephalometric analysis.[11-13] The diagnostic value of analysis depends on the accuracy and the reproducibility of landmark identification.[14-17] The most recent studies in orthodontics, however, still rely on conventional cephalometric analysis depending on human tasks.[8,18-20] A completely automated approach has thus gained attention with the aim of alleviating human error due to the analyst's subjectivity and reducing the tediousness of the task.[21-27]

Since the first introduction of an automated landmark identification method in the mid-1980s,[28] numerous methods of artificial intelligence techniques have been suggested. However, in the past, the various approaches did not seem to be accurate enough for use in clinical practice.[24] Rapidly evolving newer algorithms and increasing computational power are providing improved accuracy, reliability, and efficiency. Recent approaches for fully automated cephalometric landmark identification have shown significant improvement in accuracy and are raising expectations for daily use of these automatic techniques.[21,25,27] According to preceding research, the random forest technique was one of the most popular machine learning methods. Recently, an advanced machine learning method called "deep learning" has been receiving the spotlight.[23] However, the first step towards applying this latest method to the automated cephalometric analysis system is

just recently being taken.[21]

Currently available automated landmark detection solutions previously focused on a limited set of skeletal landmarks, less than 20, limiting its application either in determining precise anatomical structures or in providing soft-tissue information.[21,25-27] Cephalometric landmarks are not solely used for cephalometric analysis for skeletal characteristics. A much greater number of both skeletal- and soft-tissue landmarks are necessary for evaluation, treatment planning, and predicting treatment outcomes. It has repeatedly been emphasized that when a greater number of anatomic landmark locations are used, a more accurate prediction of treatment outcome will result.[19,29-32] In order to effectively apply automatic cephalometrics in clinical practice, the computational performance would also be an important factor, especially when the system has to deal with a large number of landmarks to be identified. Previous research revealed that the systems based on the random forest method detected 19 landmarks in several seconds.[27] Recently, one of the deep learning methods, You-Only-Look-Once (YOLO) has shown to require shorter time in detecting objects.[33] A comparison among the latest machine learning algorithms in terms of computational efficiency might be of interest to clinical orthodontists.

The purpose of this study was to compare accuracy and computational performance of two latest machine learning methods for automatic identification of cephalometric landmarks. This study applied two different algorithms in identifying 80 landmarks: 1) YOLO version 3 (YOLOv3) based method with modification,[33,34] and 2) the Single Shot Detector (SSD)

2

based method.[35] The null hypothesis was that there would be no difference in accuracy and computational performance between the two automated landmark identification systems.

II. MATERIALS AND METHODS

1. Subjects

A total of 1,311 lateral cephalometric radiograph images were selected and downloaded from the archive of Seoul National University Dental Hospital Picture Aided Communication System Server (INFINITT Healthcare Co., Ltd., Seoul, Korea). In later stages, 1,028 images were randomly selected as learning data, and the remaining 283 images played a role as new test data. Images of patients with growing capacity, fixed orthodontic appliances, massive dental prostheses, and/or surgical bone plates were all included. The exclusion criteria were only limited to extremely poor quality images that made landmark identification practically impossible. **Table 1** provides further information of study subjects. The institutional review board for the protection of human subjects reviewed and approved the research protocol (institutional review board numbers, S-D 2018010 & ERI 19007).

2. Manual identification of cephalometric landmarks

Out of 1,311 lateral cephalometric images, a total of 80 landmarks including 2 vertical reference points that were located on the free-hanging metal chain on the right side, 46 skeletal, and 32 soft-tissue landmarks (**Figure 1**) were manually identified by a single examiner with over 28 years of clinical orthodontic experience. A modification of a commercial cephalometric analysis software (V-Ceph version 8, Osstem Implant Co. Ltd, Seoul, Korea) was used to digitalize the records for the 80 landmarks. Among them, 27

were arbitrary landmarks to render smooth line drawing of anatomic structures, and 53 were conventional landmarks which have been well-accepted in clinical orthodontic practice (**Table 2**). A 30×30 pixel region was used as the label during the annotation process.

3. Two Deep Learning Systems

Two systems were built on a server running Ubuntu 18.04.1 LTS OS with a Tesla V100 GPU acceleration card (NVIDIA Corp., Santa Clara, CA). One system was based on YOLOv3,[34] the other one was based on SSD.[35] Learning data (N = 1,028) trained the two machine learning algorithms. Manually recorded location data of 80 landmarks served as standardized inputs in this learning process.

The target image was resized to 608×608 pixels from the original size of 1,670×2,010 pixels for optimal deep learning. One millimeter was equal to 6.7 pixels. While learning, each image along with its corresponding landmark labels was then passed through convolutional neural network (CNN) architecture for both YOLOv3 and SSD. When a previously unseen image was tested, the trained algorithms would automatically find each landmark with the highest probability through 3 different detections. Through the process of merging and selection of the highest probability of a location, most of the 80 landmarks were successfully identified. When the system failed to detect the most probable point for a landmark, the supplementary function used the pre-recorded relative coordinate information to automatically identify the missing landmark. The process was applied equally for both algorithms.

4. Test Procedures and Comparisons between the two systems

To test accuracy and computational efficiency between the two systems, 283 test data that were not included in the learning data were used.

The accuracy of the two systems are reported as point-to-point errors that were calculated as the absolute distance value between the reference position and the corresponding automatically identified landmarks.

To visualize and evaluate errors, 2-dimensional scattergrams and 95% confidence ellipses based on chi-square distribution[36-38] for each landmark were depicted.

To follow the format of previous accuracy reports, thereby making analogous comparisons with previous results possible, the successful detection rates (SDR) for 2, 2.5-, 3, and 4 mm ranges were calculated for 19 landmarks that were previously utilized in the literature.[21]

Computational performances were reported as the mean running time required to identify 80 landmarks of an image under this study's laboratory conditions.

Figure 2 summarizes the overall experimental design of the current investigation.

The differences in the test errors between YOLOv3 and SSD were compared by the *t*-test at the provability of 0.05 with the Bonferroni correction of alpha errors. All of the statistical analyses were performed by Language R (Vienna, Austria).[39]

Ⅲ. RESULTS

YOLOv3 algorithm outperformed SSD in accuracy for 38/80 landmarks. The other 42/80 landmarks did not show statistically significant difference between the two methods. None of the landmarks were found to be more accurately identified by the SSD method (**Table 3 and Figure 3**).

When compared with the top benchmark in the literature to date so far,[21] YOLOv3 showed approximately 5% higher SDR in all ranges (**Figure 4**).

The error scattergams revealed that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did. Among the scattergrams, some landmarks were detected by YOLOv3 with smaller error range (**Figure 5 C, Q, R**). Some landmarks were detected by YOLOv3 with less biased tendency (**Figure 5 E, F, H, I, J, K**). Some landmarks revealed to have similar error distribution (**Figure 5 O, P**). However, most of the figures show YOLOv3 has not only smaller ellipses in size but also a more homogenous distribution of detecting errors irrespective of the direction (**Figure 5 A, B, D, G, L, M, N**). The latter can be seen by a more circular shape of the ellipses of YOLOv3, while SSD has crushed-shaped ellipses.

The mean time spent in identification and visualization of the 80 landmarks per an image was recorded as 0.05 and 2.89 seconds for YOLOv3 and SSD, respectively.

IV. **DISCUSSION**

The present study was performed to investigate which kind of latest deep learning method would produce the most accurate results in automatically identifying cephalometric landmarks. Applying artificial intelligence techniques to routine clinical procedures is gaining global attention in medical fields.[1-3] Among these, automatic cephalometric landmark identification and analysis is the most popular topic in orthodontics. Although 3-dimensional images have gained popularity these days,[6,40-43] 2-dimensional cephalometric analysis is still a vital tool in orthodontic diagnosis and treatment planning since it provides information regarding a patient's skeletal- and soft tissue. Nevertheless, until the mid-2000s, the developed algorithms did not seem accurate enough for clinical purposes.[24] More recently, annual global competitions revealed impressive improvements in the accuracy of automated cephalometric landmark identification.[21,26,27] In fact, recent approaches based on algorithms showed accuracy comparable to an experienced orthodontist.[25,27] The result of the present study demonstrated that YOLOv3 was better than SSD. Furthermore, the accuracy results of the present study showed that YOLOv3 was better than other top benchmarks to date so far.[21,26,27] Among the previous literature, the most accurate result was produced after applying convolutional neural networks (CNN) which identified 19 landmarks.[21] The present study identified significantly more, 80 landmarks that can readily be extrapolated for clinical use in predicting treatment outcomes.[29-32] For clinical purposes, data from cephalometric landmark identification could readily be extended even to predict and visualize soft tissue changes after the treatment. For the aforementioned purposes, the previous international competitions

dealing with 19 landmarks might not meet the clinical needs in orthodontic practice.[26,27]

Since the random forest method was first introduced,[44] the algorithm has become one of the most popular for developing automatic cephalometric landmark identification systems.[25-27,45,46] During developing a prediction algorithm, there could be an issue that the algorithm fits well in the training data, but poor in the new testing data. This phenomenon is known as overfitting.[47] While using the random forest method has advantages, such as less overfitting issues, there also exists a limitation: it is difficult to predict a response value out of the range of the training set and it is sensitive to the image quality and size.[44,48] The origin of the deep learning method dates back to 1980s.[49] Back then, the overall computational performance was so poor that its application to daily life was not possible. However, continuous developments in software and hardware made the technology evolving.[50,51] In 2012, among those deep learning methods, the convolutional neural network model (CNN) showed outstanding performance in an image classification task.[52] Consequently, applications of deep learning models to overall technology are becoming reality.[23] Papers focusing on one of them, CNN, have been rapidly accumulating.[1,4,21] Regarding automated cephalometric landmark identification, efforts to apply CNN have begun relatively recently.[21] Therefore, application and comparison of latest deep learning models in automatic landmark identification system might be a valuable addition to our knowledge base.

In 2016, with the aim of real-time object detection in testing images, two novel algorithms came out, namely YOLO and SSD.[33,35] While both of them had CNN architectures, additional simplifying mechanisms shortened the detecting process and

outperformed region-based convolutional neural network (R-CNN) based methods in computational performance. YOLO uses CNN to reduce the spatial dimension detection box. It performs a linear regression to make boundary box prediction. The purported advantage of YOLO is fast computation and generalization. YOLO based machines can detect objects on artwork samples even if they were trained by a natural image set. We expected this same characteristic to be advantageous for automated cephalometric analysis when applied to radiographs with various image size and quality. YOLO needs GPUs, but due to its inherent ability to extract features automatically for learning, it is very robust.[33,34]

The accuracy measured by point-to-point errors showed that none of the landmarks was identified more accurately by SSD than by YOLOv3. On the other hand, the SSD system revealed to have accuracy inferior to pre-existing state-of-the-art works.[21,27] In the case of SSD, the size of the detecting box is usually fixed and used for simultaneous size detection. Therefore, the purported advantage of SSD is known to be the simultaneous detection of objects with various sizes. However, in landmark identification of cephalometric radiographs, the size of the detecting box is generally fixed. This was conjectured to be one reason for the poorer detection performance of SSD. A well-known limitation of both YOLO and SSD was that their accuracy was inferior to other methods when the size of objects is small. However, in the latest version of YOLO (YOLOv3) claimed to improve its accuracy to the level of other pre-existing methods while keeping the aforementioned advantages.[34]

Some of the landmarks are prone to error in the vertical direction while others show

greater errors in the horizontal direction.[24,53] Hence evaluating the accuracy based only on the linear distance might not be informative enough. Therefore, 2-dimensional scattergrams and 95% confidence ellipses of 80 landmarks were depicted. As shown in **Figure 5 A~R**, YOLOv3 revealed to have ellipses with smaller sizes and more circular shapes. In other words, YOLOv3 was not just more accurate but also resulted in a more isotropic shape of error patterns than SSD. This feature might be another advantage of YOLOv3.

YOLOv3 showed approximately 5% higher accuracy compared with the top benchmarks in the literature.[21] There could be an argument that the testing set was different from the pre-existing research. In this study, however, the test images were selected from patients who had severe type of either mandibular deficiency, prognathism, or facial asymmetry. They had undergone orthognathic surgeries eventually. The descriptive summary in **Table 1** reflect and match well with the current trend of patients seeking a university affiliated dental healthcare institution that has a high proportion of orthodontic patients with severe skeletal discrepancies.[54,55] Even with a more difficult condition rather than good looking subjects, the result seemed satisfactory.

The computational time of an automated cephalometric landmark identification system might be a concern to clinicians. The mean time spent per image was 0.05 seconds for YOLOv3 and 2.89 seconds for SSD under this study's laboratory conditions. Even with an extensive number of landmarks to be identified, both algorithms showed excellent speed. Based on these evaluations, YOLOv3 seemed to be promising as a fully automated cephalometric landmark identification system for use in clinical practice.

The application of artificial intelligence in automated cephalometric landmark identification has gained global attention and is certainly not confined to orthodontics only. Machine learning systems may lessen the burden of the clinician and alleviate human errors in cephalometric landmark detection and reduce the time required for preparing orthodontic diagnosis. By gathering radiographic data automatically, the YOLOv3 method may also help reduce human tasks and the time required for both research and clinical purposes.

One strength of the present study is that the data included the largest number of learning (n = 1,028) and test data (n = 283) ever investigated. The number of cephalometric landmarks was also the greatest, 80 landmarks that included those on the soft tissue from glabella to the terminal point on the neck.

VI. CONCLUSIONS

1. YOLOv3 showed higher accuracy in automated cephalometric landmark identification and the accuracy was approximately 5% higher accuracy compared with top benchmarks in the literature.

2. YOLOv3 outperformed SSD in the accuracy and computational time.

3. YOLOv3 also demonstrated a more isotropic form of detection errors than SSD did.

4. Between the two algorithms applied, YOLOv3 seemed to be a promising method for used as an automated cephalometric landmark identification system.

REFERENCES

1. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.

2. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O. Deep learning with convolutional neural network in radiology. Jpn J Radiol 2018;36:257-272.

3. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol 2017;18:570-584.

4. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 2018;77:106-111.

5. Jung SK, Kim TW. New approach for the diagnosis of extractions with neural network machine learning. Am J Orthod Dentofacial Orthop 2016;149:127-133.

6. Sam A, Currie K, Oh H, Flores-Mir C, Lagravere-Vich M. Reliability of different three-dimensional cephalometric landmarks in cone-beam computed tomography : A systematic review. Angle Orthod 2019;89:317-332.

7. Noll D, Mahon B, Shroff B, Carrico C, Lindauer SJ. Twitter analysis of the orthodontic patient experience with braces vs Invisalign. Angle Orthod 2017;87:377-383.

8. Rossato PH, Fernandes TMF, Urnau FDA, de Castro AC, Conti F, de Almeida RR et al. Dentoalveolar effects produced by different appliances on early treatment of anterior open bite: A randomized clinical trial. Angle Orthod 2018;88:684-691.

9. Montúfar J, Romero M, Scougall-Vilchis RJ. Automatic 3-dimensional cephalometric landmarking based on active shape models in related projections. Am J Orthod Dentofacial Orthop 2018;153:449-458.

10. Montúfar J, Romero M, Scougall-Vilchis RJ. Hybrid approach for automatic

cephalometric landmark annotation on cone-beam computed tomography volumes. Am J Orthod Dentofacial Orthop 2018;154:140-150.

11. Baumrind S, Frantz RC. The reliability of head film measurements. 1. Landmark identification. Am J Orthod 1971;60:111-127.

12. Houston WJ. The analysis of errors in orthodontic measurements. Am J Orthod 1983;83:382-390.

13. Houston WJB, Maher RE, McElroy D, Sherriff M. Sources of error in measurements from cephalometric radiographs. Eur J Orthod 1986;8:149-151.

14. Kazandjian S, Kiliaridis S, Mavropoulos A. Validity and reliability of a new edge-based computerized method for identification of cephalometric landmarks. Angle Orthod 2006;76:619-624.

15. Rudolph DJ, Sinclair PM, Coggins JM. Automatic computerized radiographic identification of cephalometric landmarks. Am J Orthod Dentofacial Orthop 1998;113:173-179.

16. Stabrun AE, Danielsen K. Precision in cephalometric landmark identification. Eur J Orthod 1982;4:185-196.

17. Kamoen A, Dermaut L, Verbeeck R. The clinical significance of error measurement in the interpretation of treatment results. Eur J Orthod 2001;23:569-578.

18. Garnett BS, Mahood K, Nguyen M, Al-Khateeb A, Liu S, Boyd R et al. Cephalometric comparison of adult anterior open bite treatment using clear aligners and fixed appliances. Angle Orthod 2019;89:3-9.

19. Kang TJ, Eo SH, Cho H, Donatelli RE, Lee SJ. A sparse principal component analysis of Class III malocclusions. Angle Orthod 2019. Available at:

http://dx.doi.org/10.2319/100518-717.1

20. Myrlund R, Keski-Nisula K, Kerosuo H. Stability of orthodontic treatment outcomes after 1-year treatment with the eruption guidance appliance in the early mixed dentition: A follow-up study. Angle Orthod 2019;89:206-213.

21. Arik SO, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham) 2017;4:014501.

22. Hutton TJ, Cunningham S, Hammond P. An evaluation of active shape models for the automatic identification of cephalometric landmarks. Eur J Orthod 2000;22:499-508.

23. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-444.

24. Leonardi R, Giordano D, Maiorana F, Spampinato C. Automatic cephalometric analysis. Angle Orthod 2008;78:145-151.

25. Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF. Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms. Sci Rep 2016;6:33581.

26. Wang CW, Huang CT, Hsieh MC, Li CH, Chang SW, Li WC et al. Evaluation and Comparison of Anatomical Landmark Detection Methods for Cephalometric X-Ray Images: A Grand Challenge. IEEE Trans Med Imaging 2015;34:1890-1900.

27. Wang CW, Huang CT, Lee JH, Li CH, Chang SW, Siao MJ et al. A benchmark for comparison of dental radiography analysis algorithms. Med Image Anal 2016;31:63-76.

28. Cohen AM, Ip HH, Linney AD. A preliminary study of computer recognition and identification of skeletal landmarks as a new method of cephalometric analysis. Br J Orthod 1984;11:143-154.

29. Suh HY, Lee SJ, Lee YS, Donatelli RE, Wheeler TT, Kim SH et al. A More Accurate

Method of Predicting Soft Tissue Changes After Mandibular Setback Surgery. J Oral Maxillofac Surg 2012;70:e553-e562.

30. Lee HJ, Suh HY, Lee YS, Lee SJ, Donatelli RE, Dolce C et al. A better statistical method of predicting postsurgery soft tissue response in Class II patients. Angle Orthod 2014;84:322-328.

31. Lee YS, Suh HY, Lee SJ, Donatelli RE. A more accurate soft-tissue prediction model for Class III 2-jaw surgeries. Am J Orthod Dentofacial Orthop 2014;146:724-733.

32. Yoon KS, Lee HJ, Lee SJ, Donatelli RE. Testing a better method of predicting postsurgery soft tissue response in Class II patients: A prospective study and validity assessment. Angle Orthod 2015;85:597-603.

33. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016.

34. Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 2018. Available at: https://arxiv.org/pdf/1804.02767.pdf.

35. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y et al. Ssd: Single shot multibox detector. Proceedings of the European conference on computer vision: Springer; 2016:21-37.

36. Donatelli RE, Lee SJ. How to report reliability in orthodontic research: Part 1. Am J Orthod Dentofacial Orthop 2013;144:156-161.

37. Donatelli RE, Lee SJ. How to report reliability in orthodontic research: Part 2. Am J Orthod Dentofacial Orthop 2013;144:315-318.

38. Donatelli RE, Lee SJ. How to test validity in orthodontic research: a mixed dentition

analysis example. Am J Orthod Dentofacial Orthop 2015;147:272-279.

39. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

40. Castillo JC, Gianneschi G, Azer D, Manosudprasit A, Haghi A, Bansal N et al. The relationship between 3D dentofacial photogrammetry measurements and traditional cephalometric measurements. Angle Orthod 2019;89:275-283.

41. Isidor S, Carlo GD, Cornelis MA, Isidor F, Cattaneo PM. Three-dimensional evaluation of changes in upper airway volume in growing skeletal Class II patients following mandibular advancement treatment with functional orthopedic appliances. Angle Orthod 2018;88:552-559.

42. Tanikawa C, Takada K. Test-retest reliability of smile tasks using three-dimensional facial topography. Angle Orthod 2018;88:319-328.

43. Feng J, Yu H, Yin Y, Yan Y, Wang Z, Bai D et al. Esthetic evaluation of facial cheek volume: A study using 3D stereophotogrammetry. Angle Orthod 2019;89:129-137.

44. Breiman L. Random Forests. Machine Learning 2001;45:5-32.

45. Mirzaalian H, Hamarneh G. Automatic globally-optimal pictorial structures with random decision forest based likelihoods for cephalometric x-ray landmark detection. Proceedings of International Symposium on Biomedical Imaging (ISBI); 2014.

46. Ibragimov B, Likar B, Pernus F, Vrtovec T. Computerized cephalometry by game theory with shape-and appearance-based landmark refinement. Proceedings of International Symposium on Biomedical imaging (ISBI); 2015.

47. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer; 2009.

48. Horning N. Random Forests: An algorithm for image classification and generation of continuous fields data sets. Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences; 2010.

49. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W et al. Backpropagation applied to handwritten zip code recognition. Neural Computation 1989;1:541-551.

50. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998;86:2278-2324.

51. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012:3642-3649.

52. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012:1097-1105.

53. Broch J, Slagsvold O, Rosler M. Error in landmark identification in lateral radiographic headplates. Eur J Orthod 1981;3:9-13.

54. Lee CH, Park HH, Seo BM, Lee SJ. Modern trends in Class III orthognathic treatment: A time series analysis. Angle Orthod 2017;87:269-278.

55. Lim HW, Park JH, Park HH, Lee SJ. Time series analysis of patients seeking orthodontic treatment at Seoul National University Dental Hospital over the past decade. Korean J Orthod 2017;47:298-305.

Table 1.   Descriptive Summary of Study Data

Study Variables		N (%)
Learning data		1,028 (100%)
Gender	Female	507 (49.3%)
Skeletal classification	Class II	178 (17.3%)
	Class III	719 (70.0%)
Test data		283 (100%)
Gender	Female	146 (51.6%)
Skeletal classification	Class II	32 (11.3%)
	Class III	251 (88.7%)
Image quality	Good	248 (87.6%)
	Fair	13 (4.6%)
	Poor	22 (7.8%)
Fixed Orthodontic appliances	Yes	140 (49.5%)

Table 2. List of 80 Cephalometric Landmarks Identified in the Present Study

Landmark number	Name	Landmark number	Name
1	Vertical reference point 1 (arbitrary)	41	Pterygoid
2	Vertical reference point 2 (arbitrary)	42	Basion
3	Sella	43	U6 crown mesial edge
4	Nasion	44	U6 mesiobuccal cusp
5	Nasal tip	45	U6 root tip
6	Porion	46	L6 crown mesial edge
7	Orbitale	47	L6 mesiobuccal cusp
8	Key ridge[a]	48	L6 root tip
9	Key ridge contour intervening point 1[a]	49	glabella
10	Key ridge contour intervening point 2[a]	50	glabella contour intervening point[a]
11	Key ridge contour intervening point 3[a]	51	nasion
12	Anterior nasal spine	52	nasion contour intervening point 1[a]
13	Posterior nasal spine	53	nasion contour intervening point 2[a]
14	Point A	54	supranasal tip
15	Point A contour intervening point[a]	55	pronasale
16	Supradentale	56	columella
17	U1 root tip	57	columella contour intervening point[a]
18	U1 incisal edge	58	subnasale
19	L1 incisal edge	59	cheekpoint
20	L1 root tip	60	point A
21	Infradentale	61	superior labial sulcus
22	Point B contour intervening point[a]	62	labiale superius
23	Point B	63	upper lip
24	Protuberance menti	64	upper lip contour intervening point[a]
25	Pogonion	65	stomion superius
26	Gnathion	66	stomion inferius
27	Menton	67	lower lip contour intervening point[a]
28	Gonion, constructed	68	lower lip
29	Mandibular body contour intervening point 1[a]	69	labiale inferius
30	Mandibular body contour intervening point 2[a]	70	inferior labial sulcus
31	Mandibular body contour intervening point 3[a]	71	point B
32	Gonion, anatomic	72	protuberance menti
33	Gonion contour intervening point 1[a]	73	pogonion
34	Gonion contour intervening point 2[a]	74	gnathion
35	Articulare	75	menton
36	Ramus contour intervening point 1[a]	76	menton contour intervening point[a]
37	Ramus contour intervening point 2[a]	77	cervical point
38	Condylion	78	cervical point contour intervening point 1[a]
39	Ramus.tip	79	cervical point contour intervening point 2[a]
40	Pterygomaxillary fissure	80	terminal point

[a] arbitrary landmarks to render smooth line drawing of anatomic structures. Landmarks #3 - #48 are skeletal landmarks and #49 - #80 are soft tissue landmakrs.

Table 3.   Comparison of Accuracy in Terms of the Point-to-Point Errors between the You-Only-Look-Once version 3 (YOLOv3) and Single Shot Detector (SSD) methods

Landmarks	YOLOv3		SSD		P value[a]	More Accurate Method
	Mean (pixel)	SD (pixel)	Mean (pixel)	SD (pixel)		
Sella	14.9	79.4	25.6	76.6	~1.0000	
Nasion	9.0	7.7	40.9	159.5	.0698	
Nasal tip	8.9	6.1	94.3	316.9	<.0001	YOLOv3
Porion	11.3	8.7	33.1	69.3	<.0001	YOLOv3
Orbitale	8.6	5.3	31.1	133.0	.3900	
ANS	15.8	13.2	366.2	497.6	<.0001	YOLOv3
PNS	9.1	7.3	29.0	86.1	<.0001	YOLOv3
Point A	13.3	9.2	244.7	458.9	<.0001	YOLOv3
U1 root tip	15.6	9.6	164.4	414.0	<.0001	YOLOv3
U1 incisal edge	7.8	4.9	396.5	584.8	<.0001	YOLOv3
L1 incisal edge	6.8	4.9	27.1	142.5	~1.0000	
L1 root tip	15.2	8.9	42.9	158.6	.2956	
Point B	14.8	10.3	83.2	260.3	.0012	YOLOv3
PM	11.9	8.9	45.1	181.6	.1912	
Pogonion	7.9	6.9	46.9	209.0	.1511	
Gnathion	8.0	5.6	50.7	259.3	.4766	
Menton	8.3	5.0	29.5	109.8	.1049	
Gonion c	15.8	8.7	88.2	111.0	<.0001	YOLOv3
Gonion a	12.7	8.4	102.9	135.9	<.0001	YOLOv3
Articulare	6.7	4.7	14.9	11.3	<.0001	YOLOv3
Condylion	11.2	8.3	22.7	56.2	.0591	
Pterygoid	13.1	36.3	18.7	49.1	~1.0000	
Basion	11.8	9.0	18.0	20.3	<.0001	YOLOv3
glabella	11.1	8.7	59.5	243.0	.0738	
nasion	10.8	7.8	54.7	208.7	<.0001	YOLOv3
supranasal tip	10.0	7.4	74.9	300.1	<.0001	YOLOv3
pronasale	7.4	5.6	40.5	187.6	.2559	
columella	9.2	7.5	33.7	103.3	<.0001	YOLOv3
subnasale	8.0	8.7	360.4	613.8	<.0001	YOLOv3
point A	10.0	6.7	71.2	311.5	.0856	
superior labial	11.4	8.9	64.7	278.0	.1133	
labium superius	9.2	6.5	46.6	211.1	.2518	
upper lip	6.3	4.6	33.3	133.7	.0619	
stomion	10.4	7.6	425.6	648.8	<.0001	YOLOv3
stomion inferius	11.0	10.1	24.4	78.5	.3807	
lower lip	5.9	3.8	86.5	132.2	<.0001	YOLOv3
labium inferius	8.3	5.9	51.0	136.5	<.0001	YOLOv3
point B	8.8	6.2	27.0	114.1	.6244	
Protuberance	10.0	8.0	39.1	172.9	.4017	
pogonion	10.6	11.8	57.7	254.9	.1648	
gnathion	16.3	15.6	35.2	91.3	.0525	
menton	13.5	14.3	52.6	250.1	.7347	

Figure 1. An image composed of the radiograph with the cephalometric landmarks used in this study. (Yellow dots) Landmark-specific information are summarized in Table 1. For the hard tissue landmarks, upper-case letters were used. For the soft tissue landmarks, lower-case letters were used.
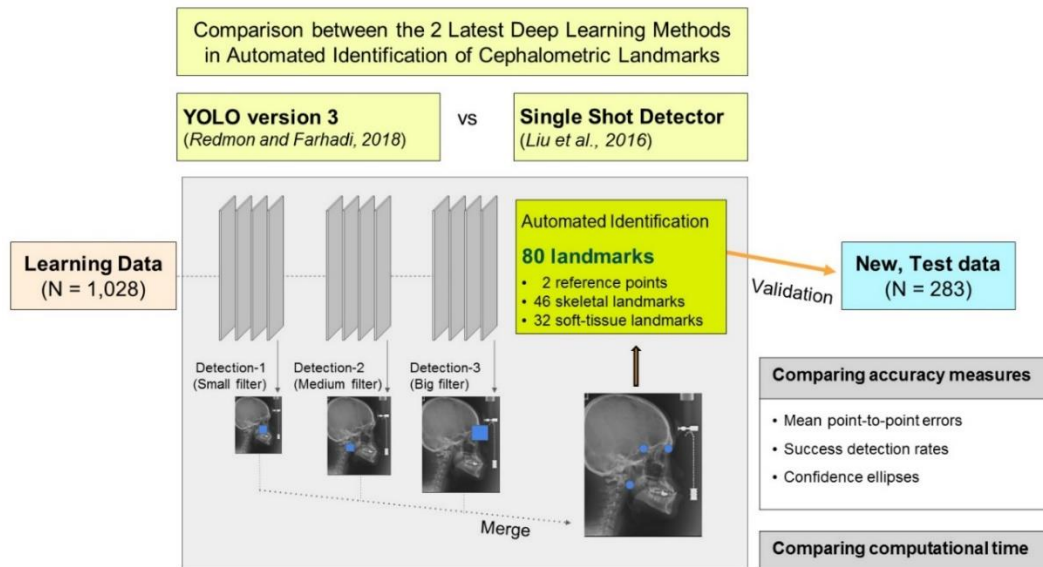
Figure 2. Diagram showing the flow of the automated landmark identification system. Each images were used for training through the convolutional neural network (CNN) architecture. The trained algorithm would automatically find each landmarks with the highest probability through 3 different detection size.
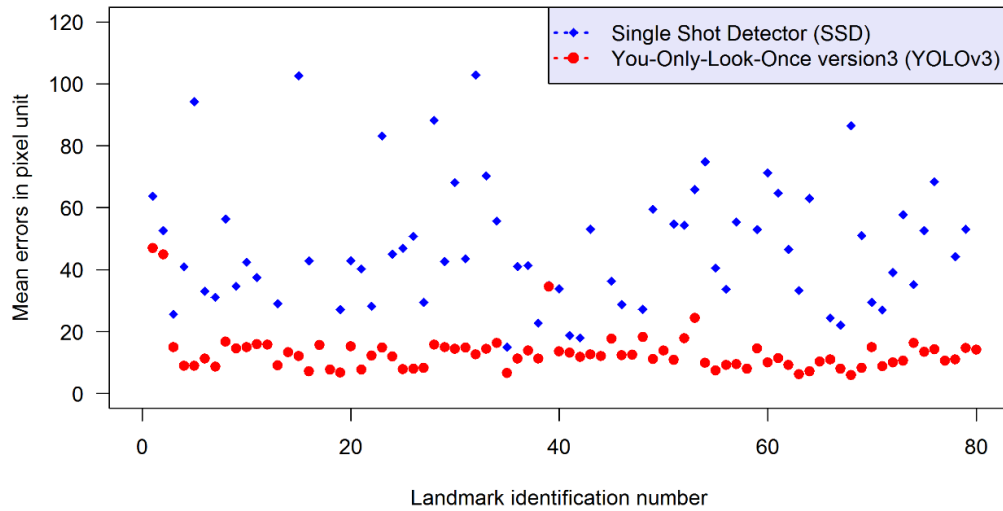
Figure 3. Point plots of landmark-specific mean point-to-point error from the You-Only-Look-Once version 3 (YOLOv3, red) and Single Shot Multibox Detector (SSD, blue) methods in automated landmark identification. All the measurements were calculated by pixel. The plot indicates that YOLOv3 was more accurate than SSD in general.
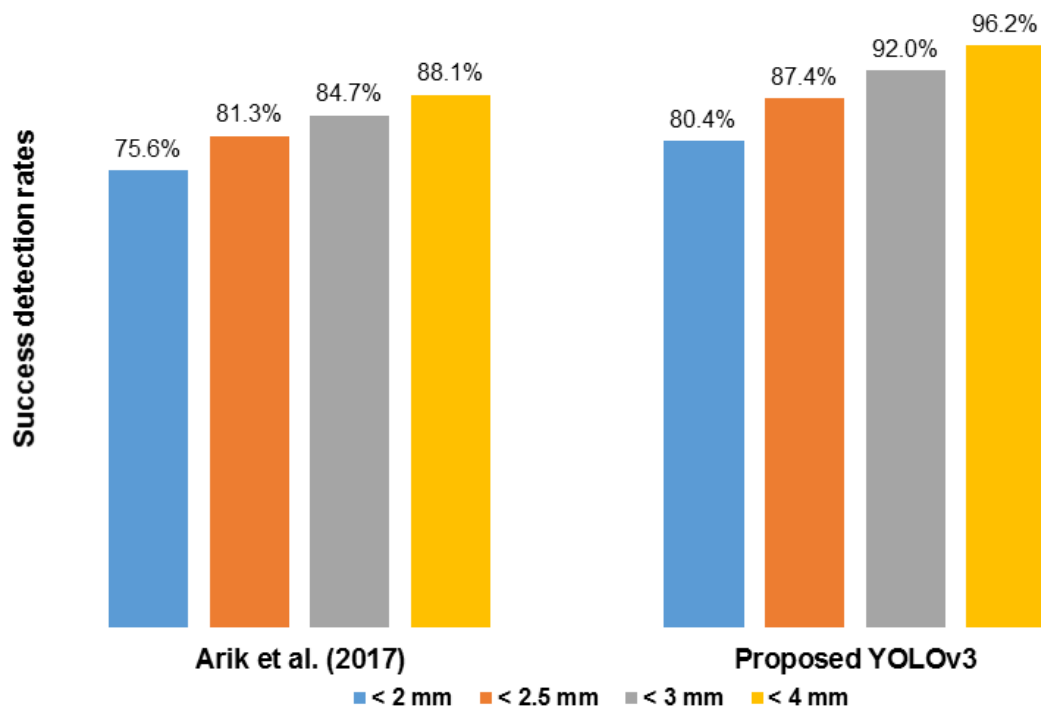
Figure 4. Success detection rates (SDR) of precision ranges, including 2 mm (blue), 2.5 mm (orange), 3 mm (gray) and 4 mm (yellow), comparing with the top accuracy results in the previous literature[21] and those from the proposed YOLOv3. The proposed YOLOv3 shows approximately 5% higher success detection rates for all ranges.
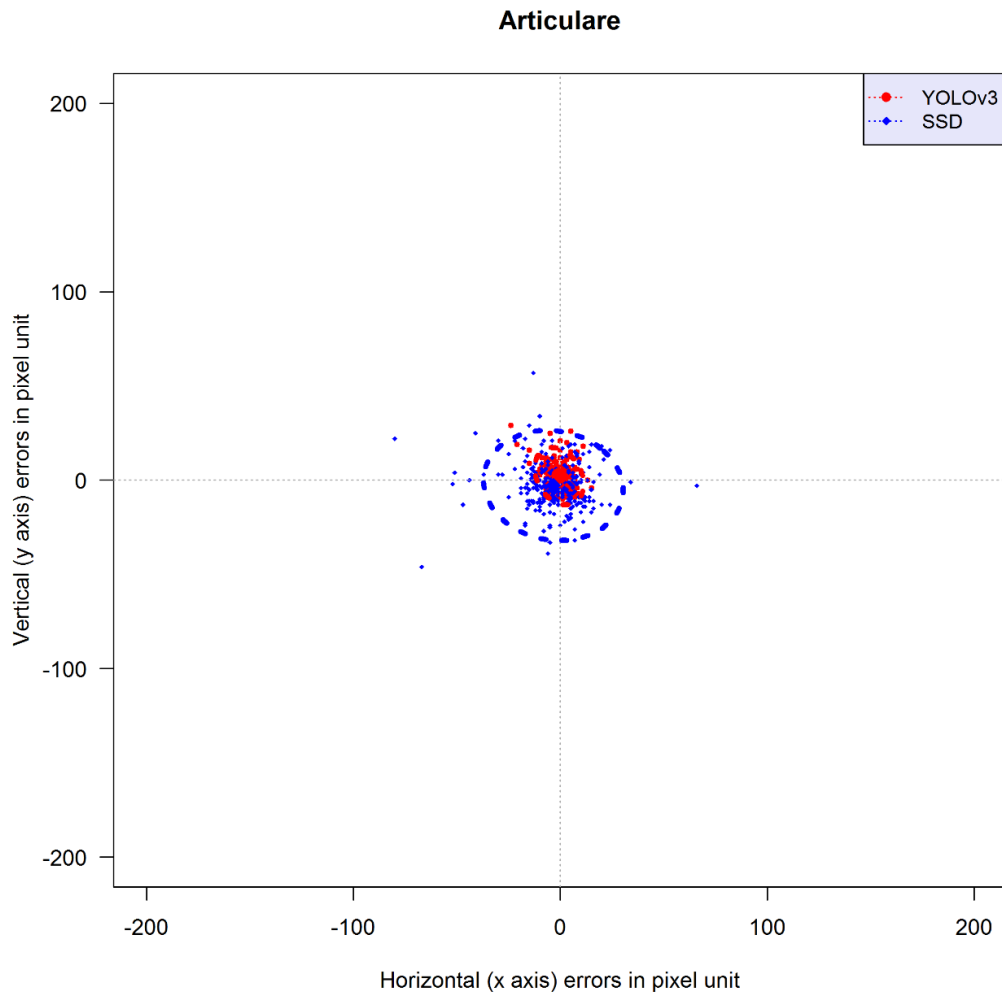
Figure 5, A. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Articulare". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did.
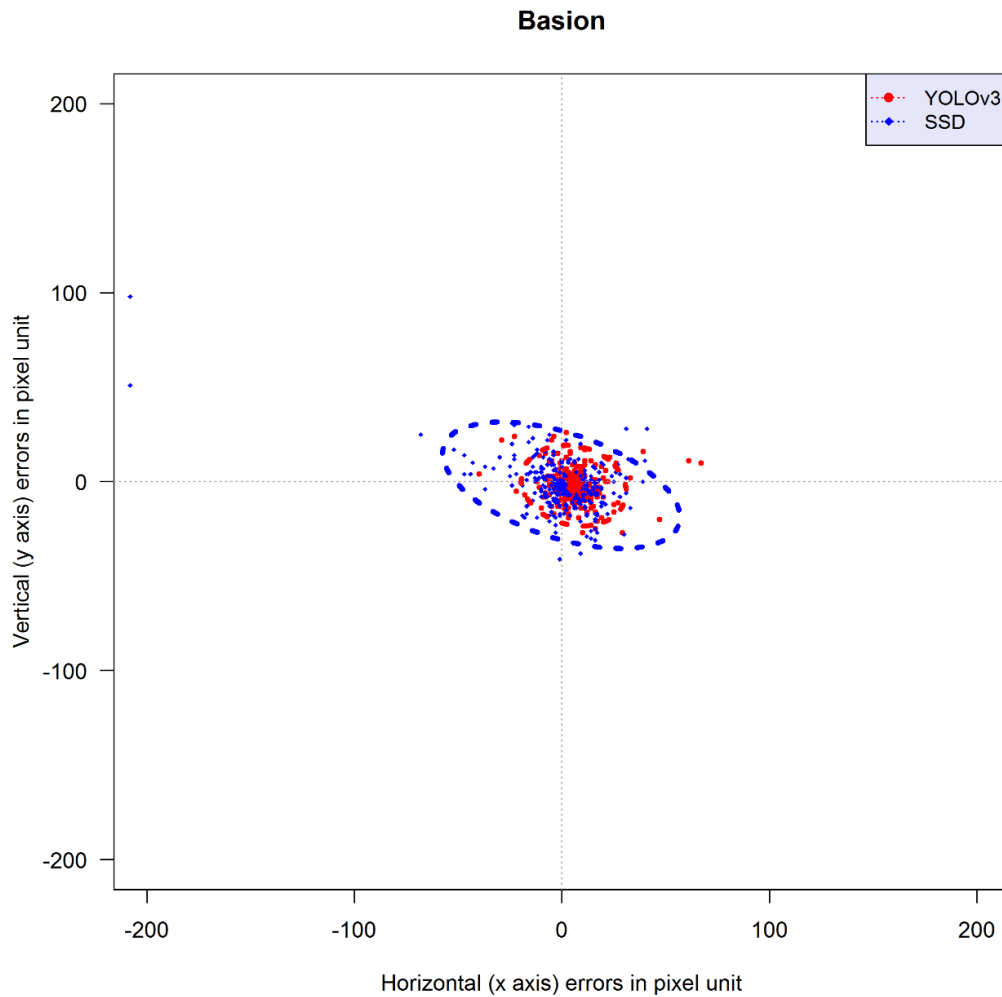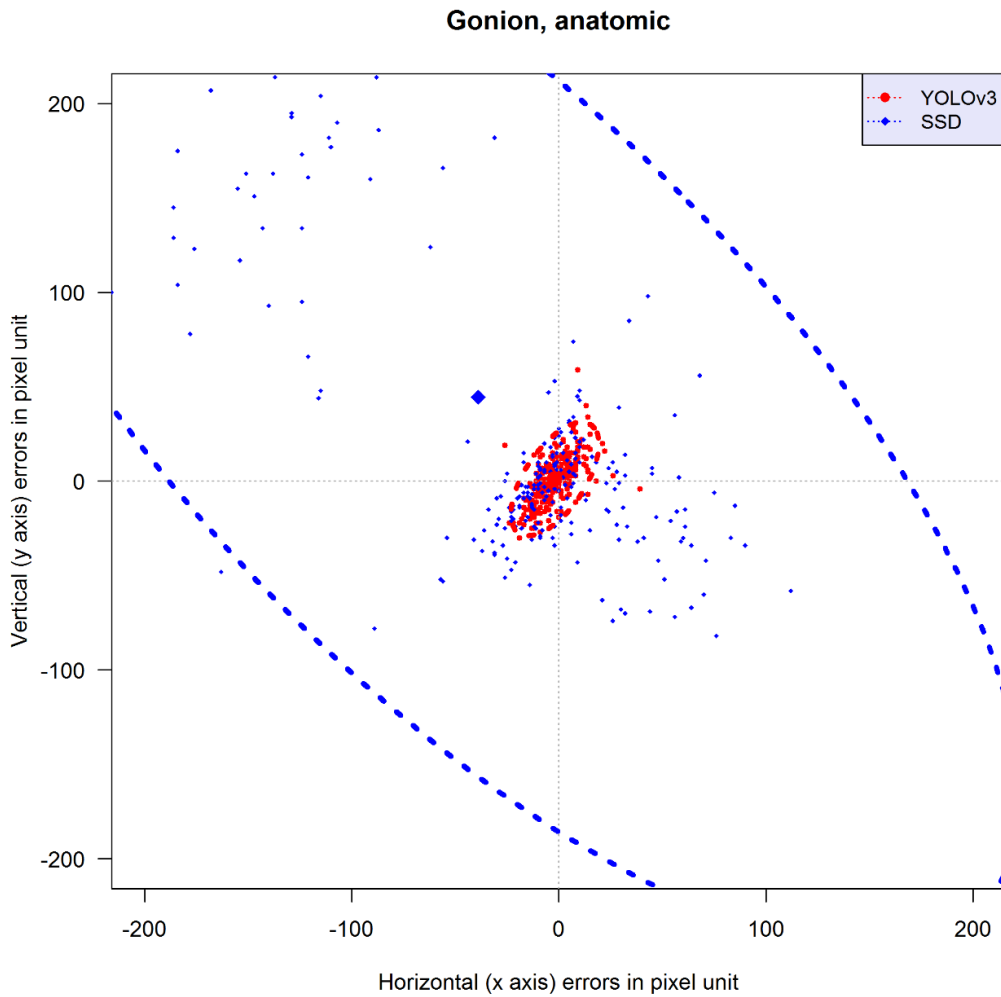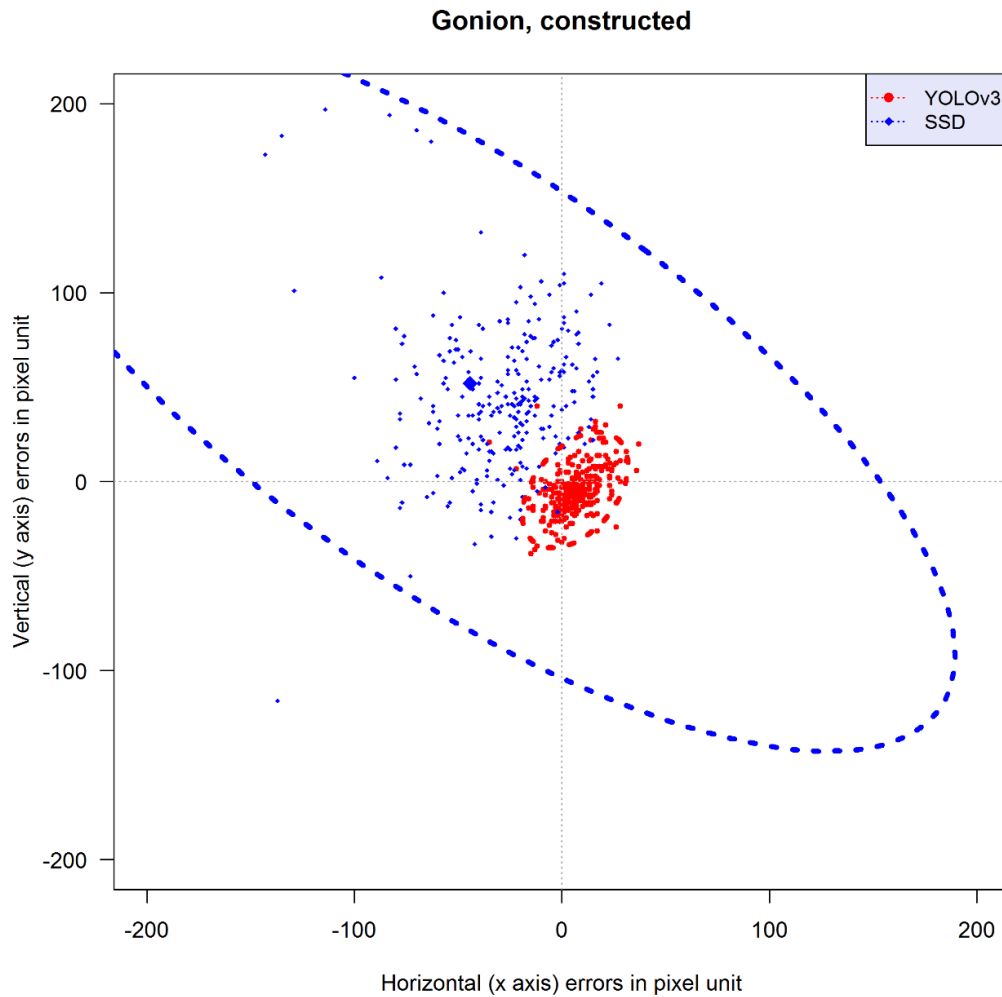
Basion

Figure 5, B. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Basion". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did.
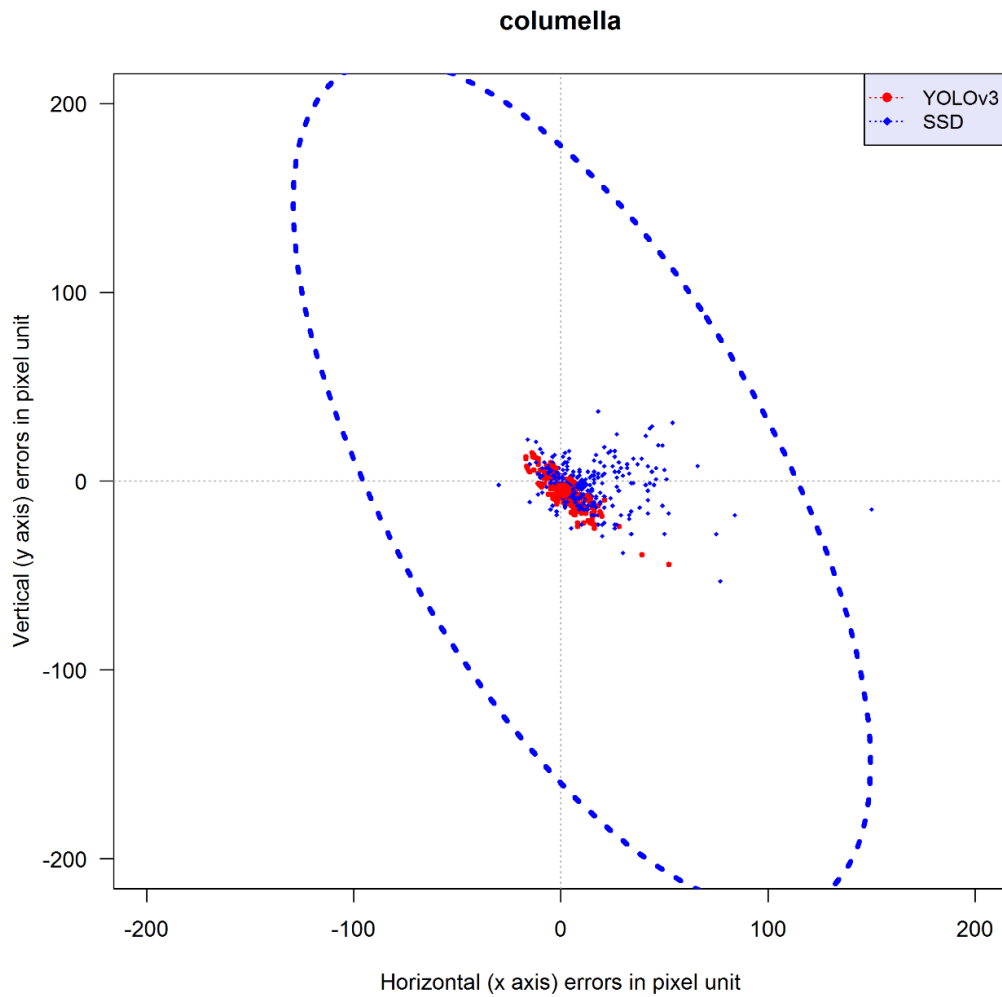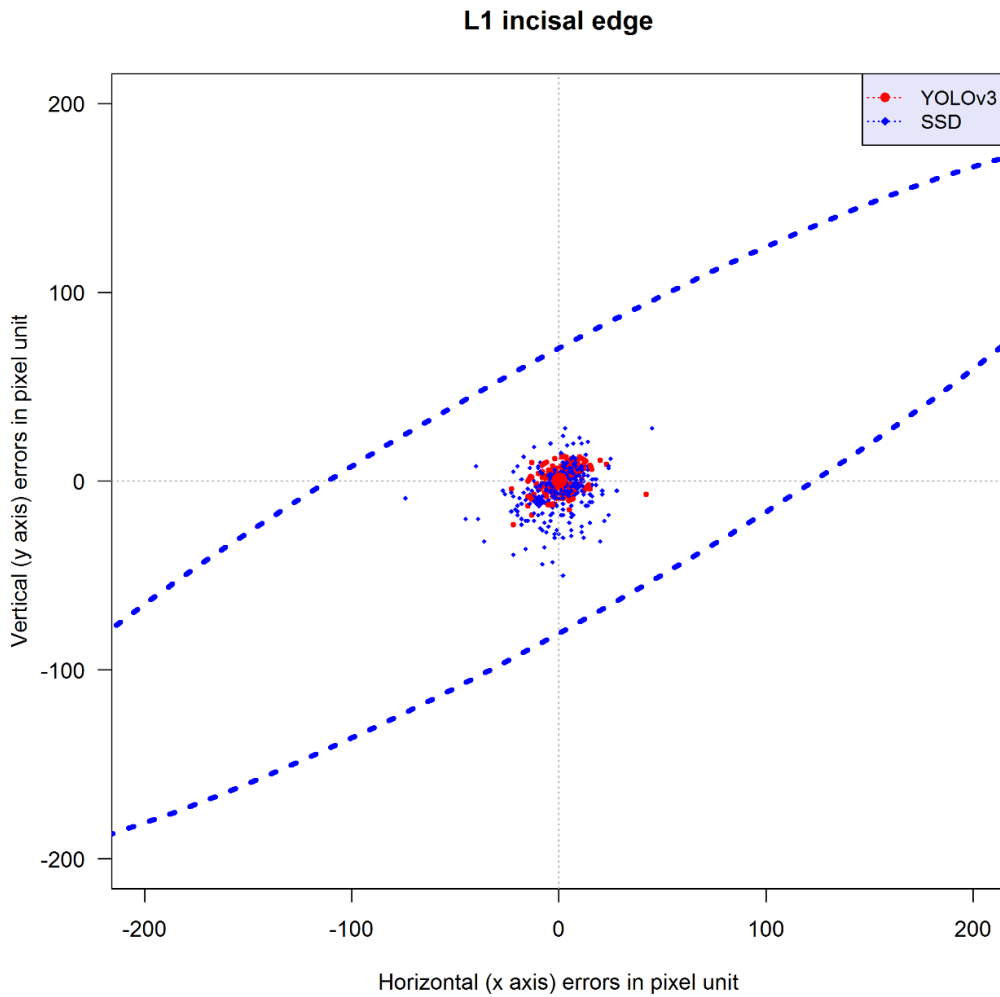
columella

Figure 5, C. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "columella". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did.
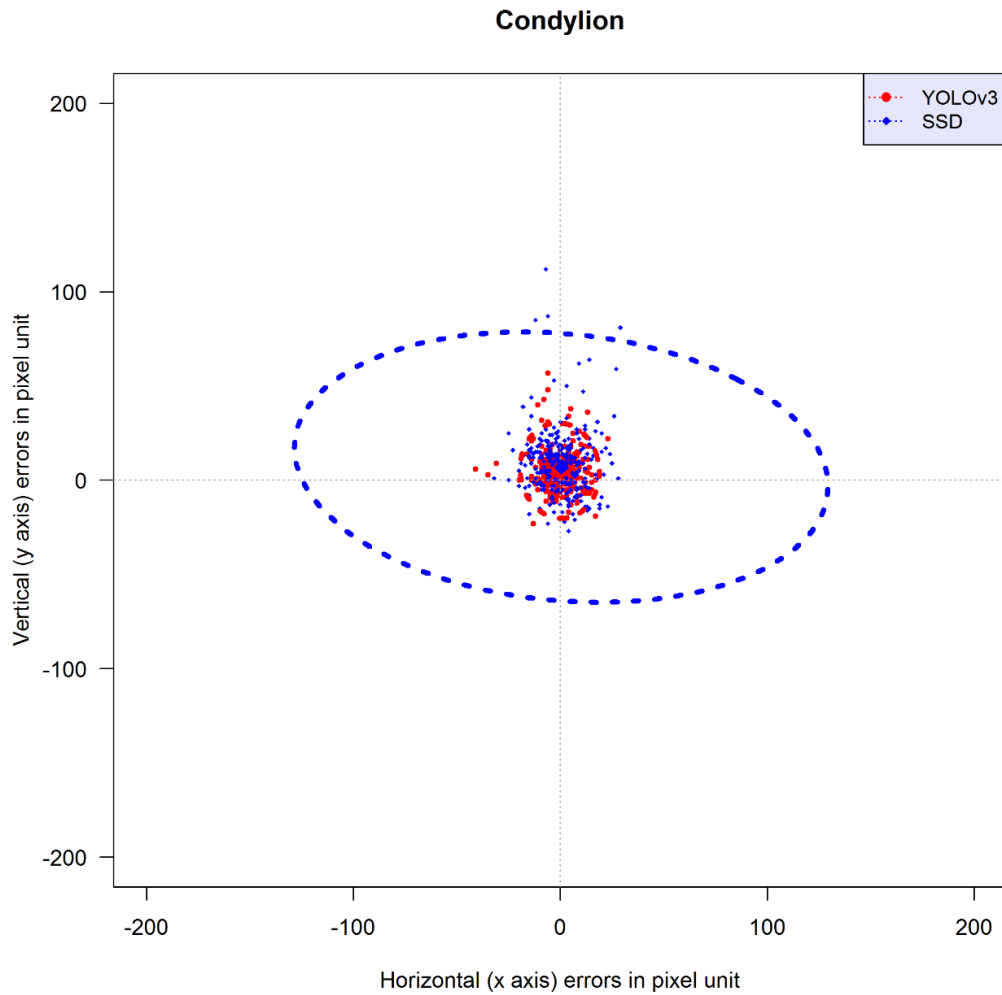
Figure 5, D. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Condylion". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did.
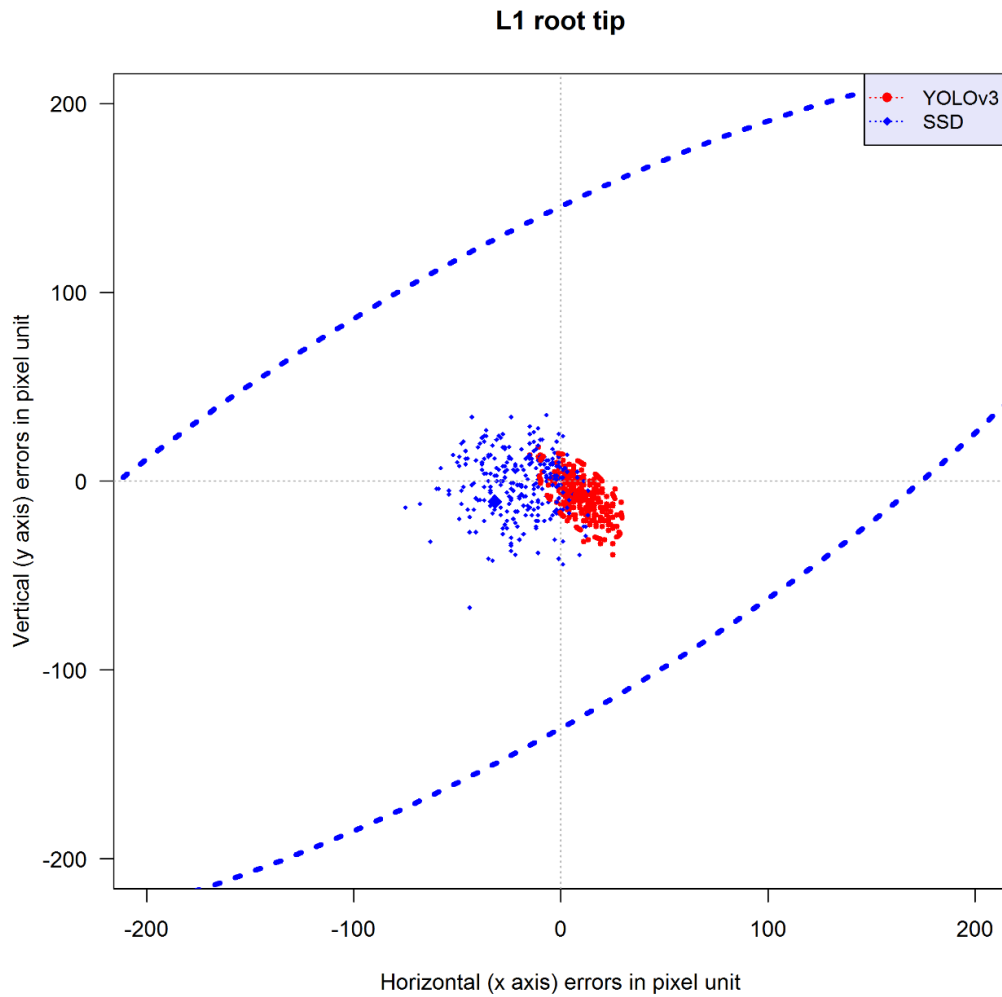
Gonion, anatomic

Figure 5, E. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Gonion, anatomic". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did. The center of the error of SSD was located distant from the origin (0,0).
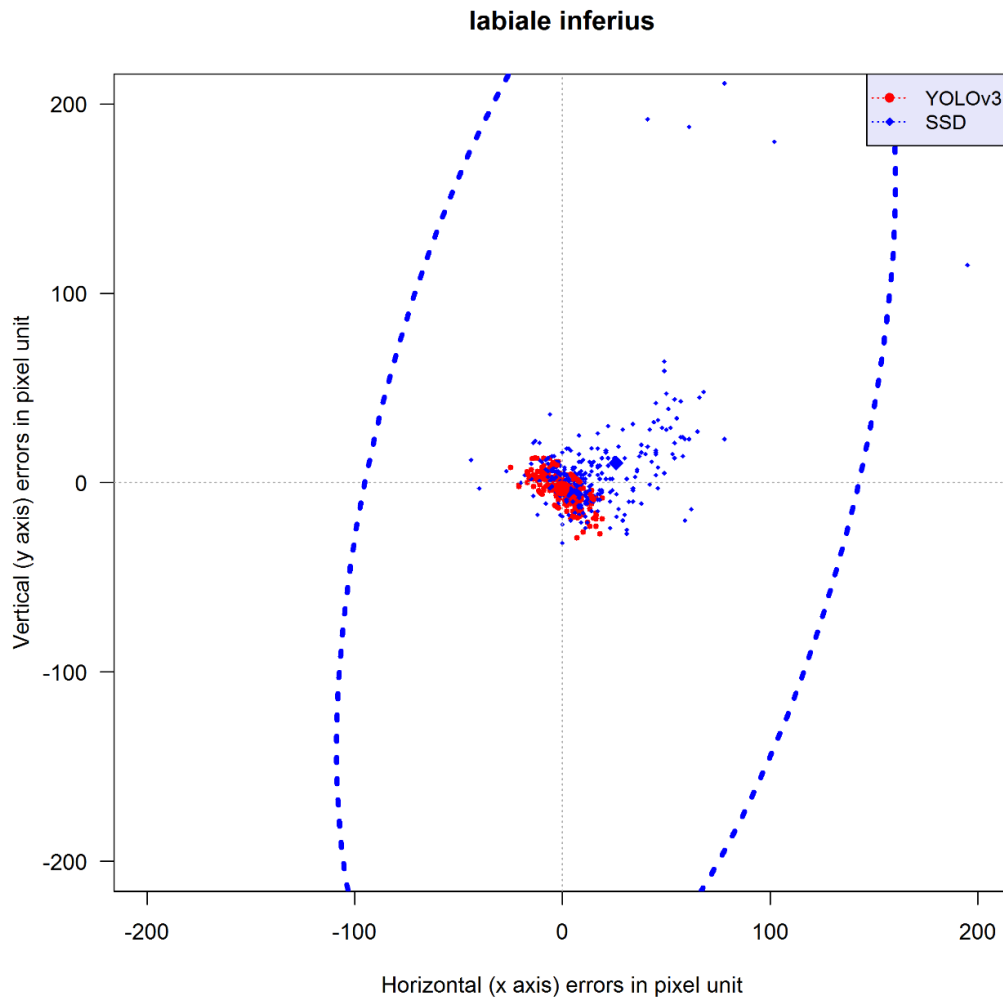
Gonion, constructed

Figure 5, F. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Gonion, constructed". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did. The center of the ellipse of SSD was located distant from the origin (0,0).
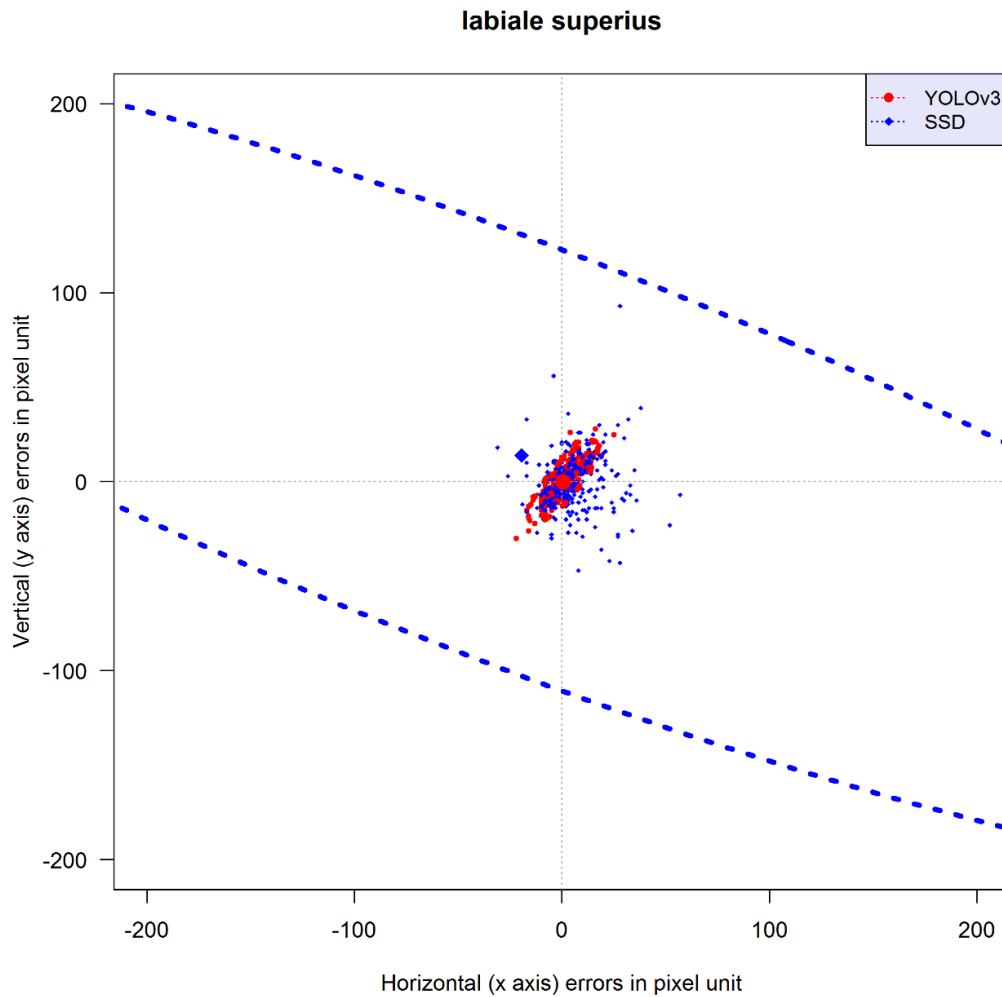
L1 incisal edge

Figure 5, G. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "L1 incisor edge". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did.

L1 root tip

Figure 5, H. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "L1 root tip". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did. The center of the ellipse of SSD was located distant from the origin (0,0).
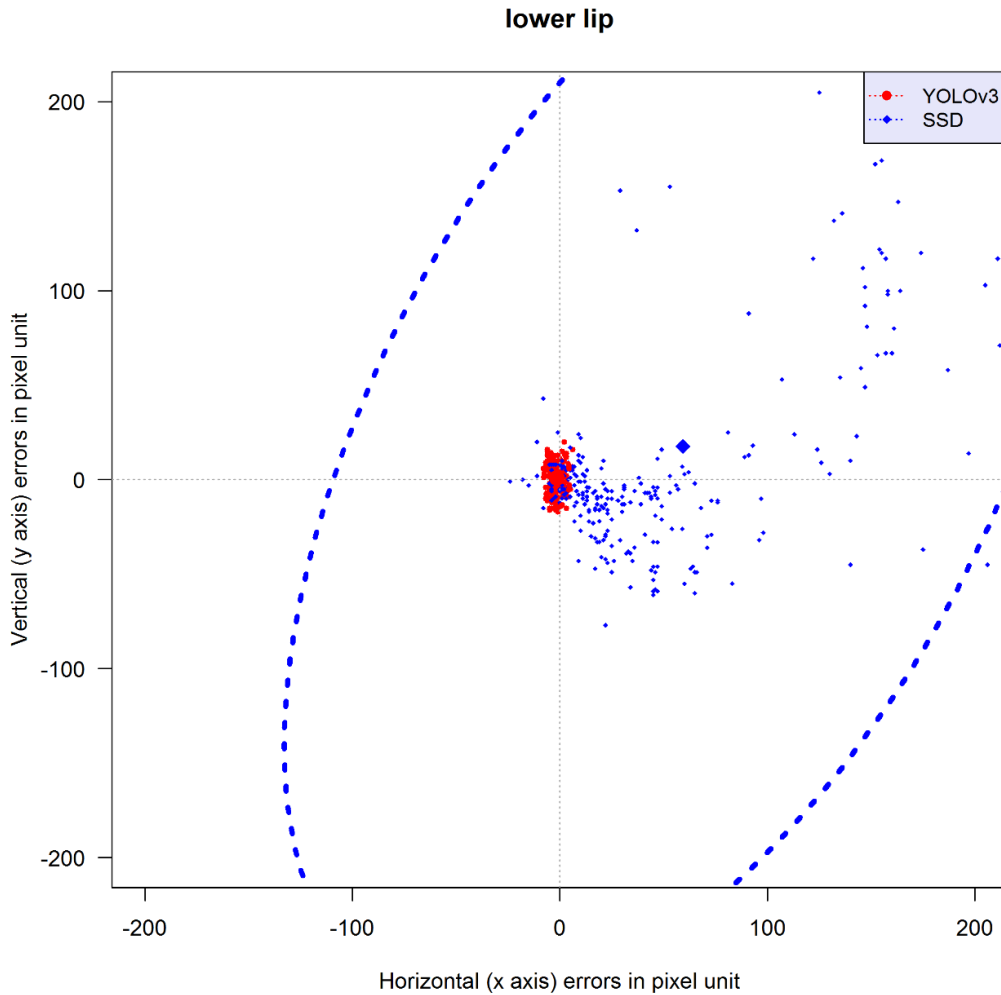
labiale inferius

Figure 5, I. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "labiale inferius". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did. The center of the ellipse of SSD was located distant from the origin (0,0).
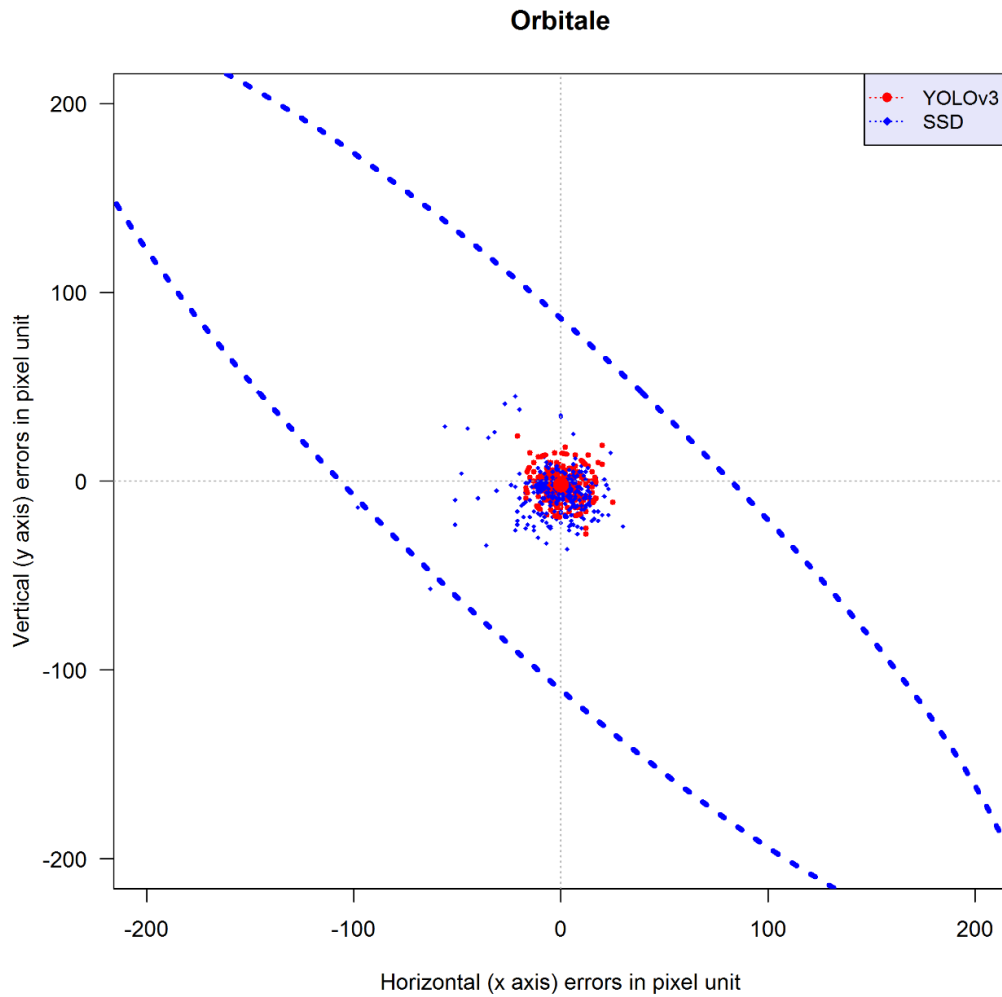
Figure 5, J. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "labiale superius". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did. The center of the ellipse of SSD was located distant from the origin (0,0).
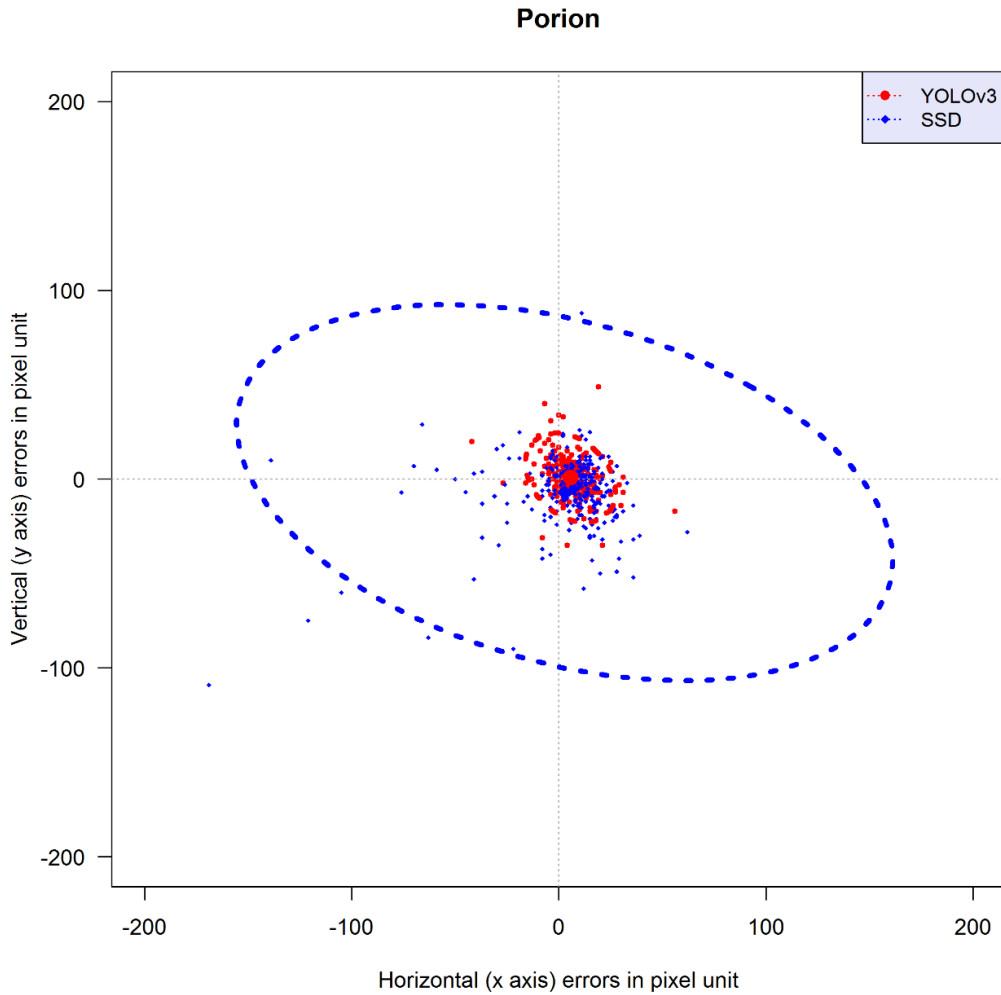
Figure 5, K. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "lower lip". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did. The center of the ellipse of SSD was located distant from the origin (0,0).

Figure 5, L. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Orbitale". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did.
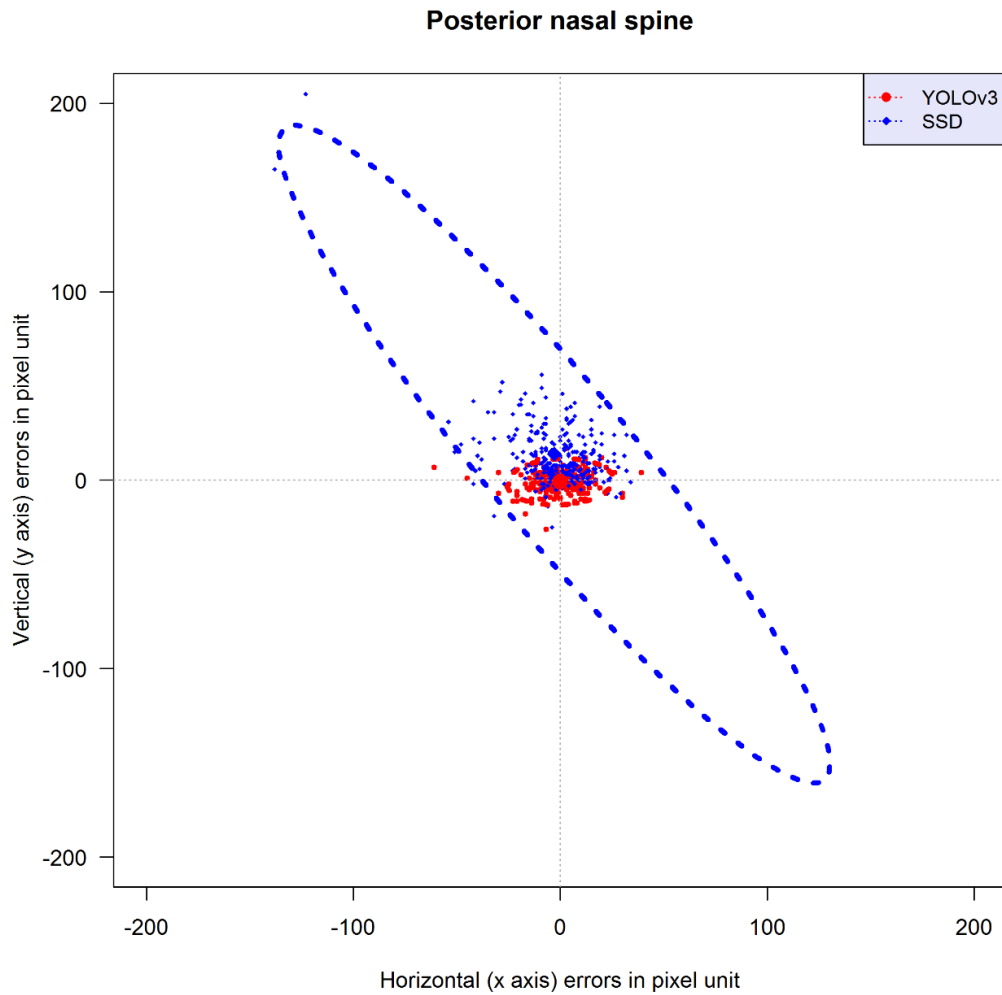
Figure 5, M. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Porion". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did.
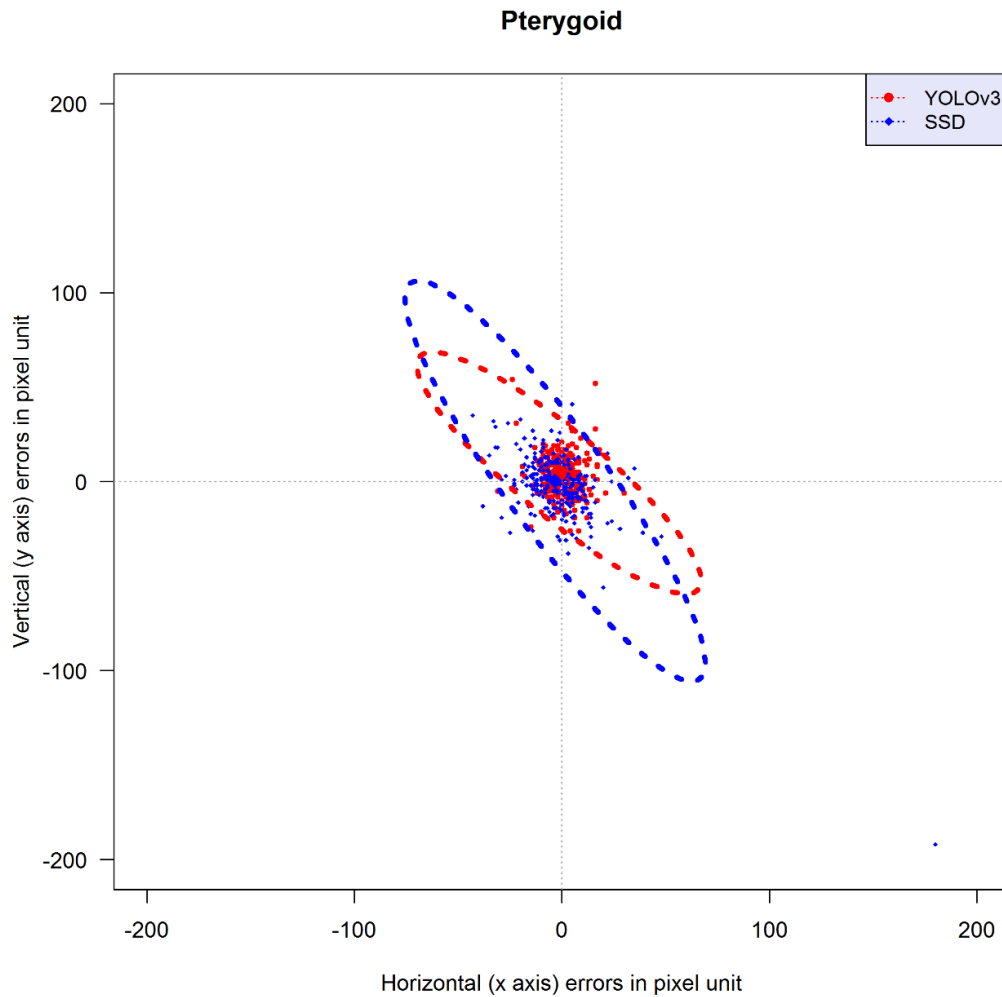
Posterior nasal spine

Figure 5, N. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Posterior nasal spine". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed not only a smaller error range, but also a more isotropic tendency than SSD did.

Figure 5, O. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Posterior nasal spine". All the measurements were calculated by pixel. The plots indicate that plots of YOLOv3 showed no significant difference compared to SSD.
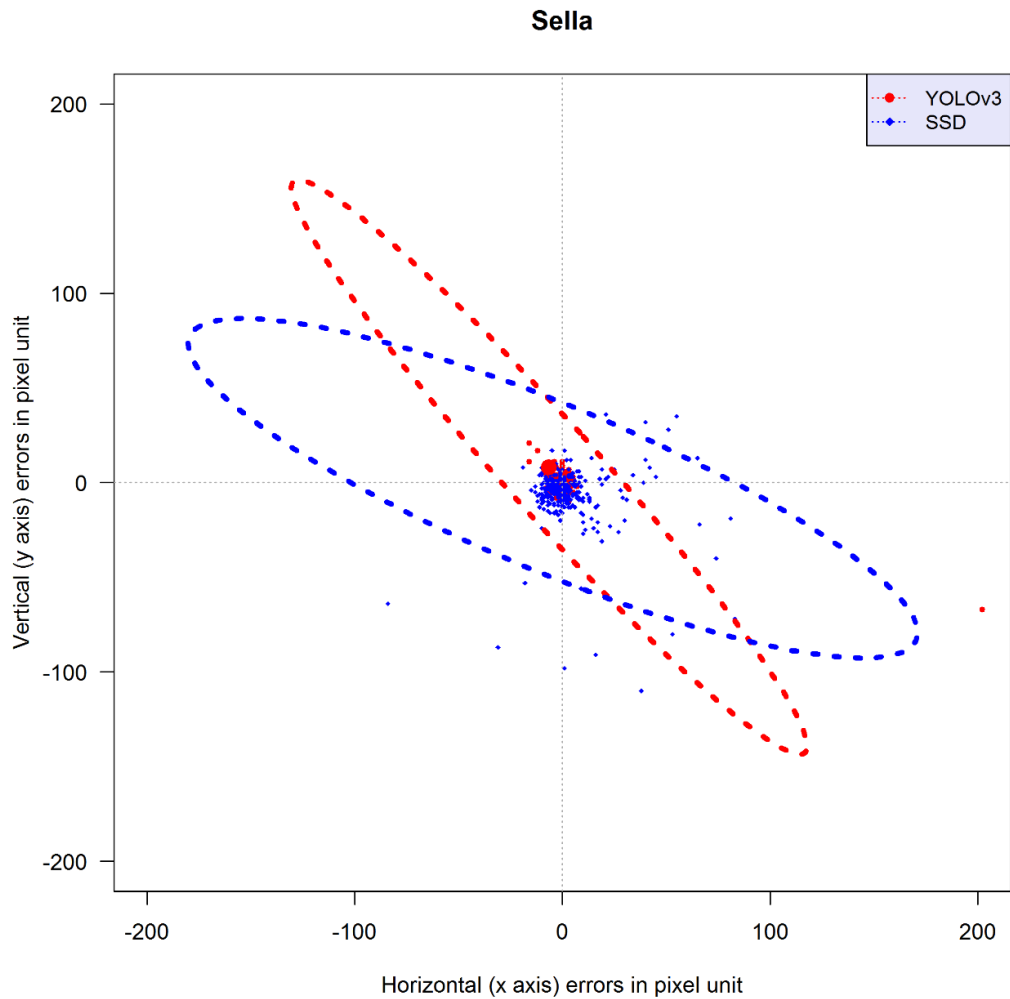
Figure 5, P. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "Sella". All the measurements were calculated by pixel. The plots indicate that plots of YOLOv3 showed no significant difference compared to SSD.
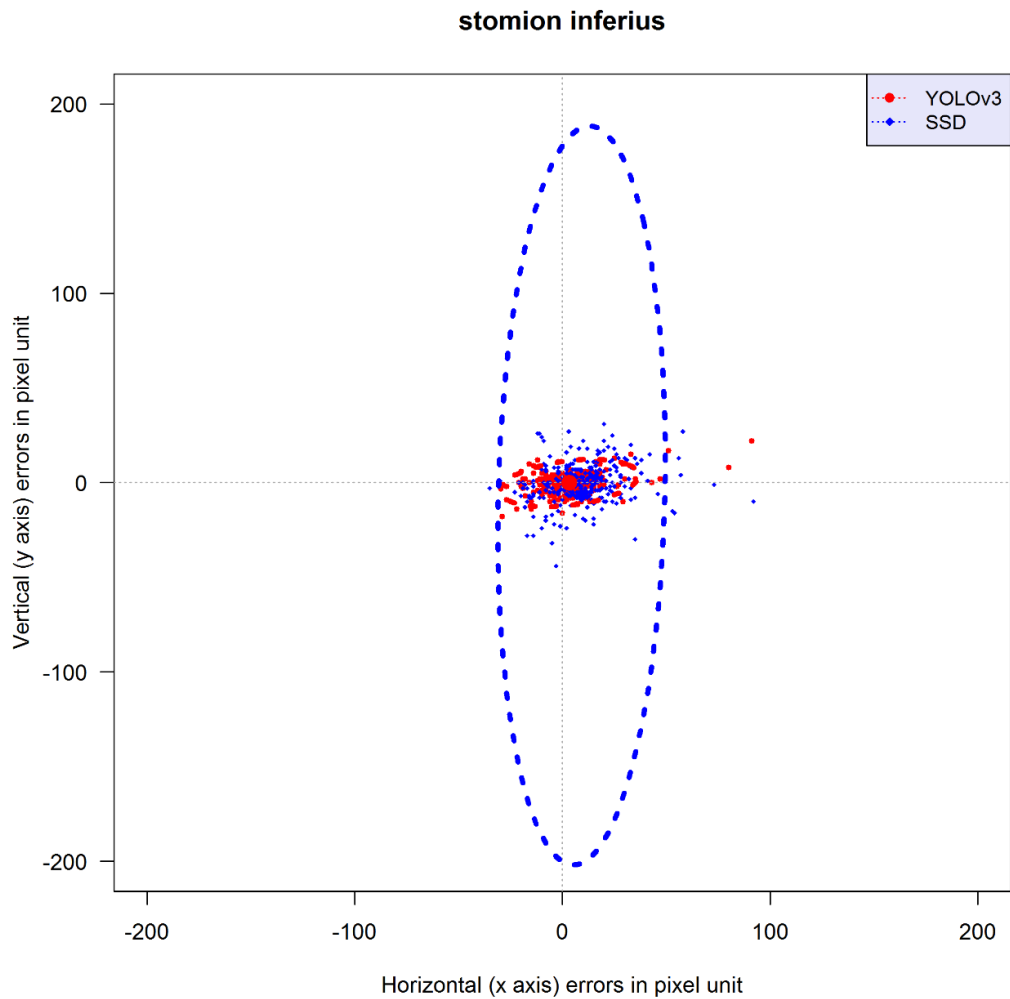
Figure 5, Q. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "stomion inferius". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did.
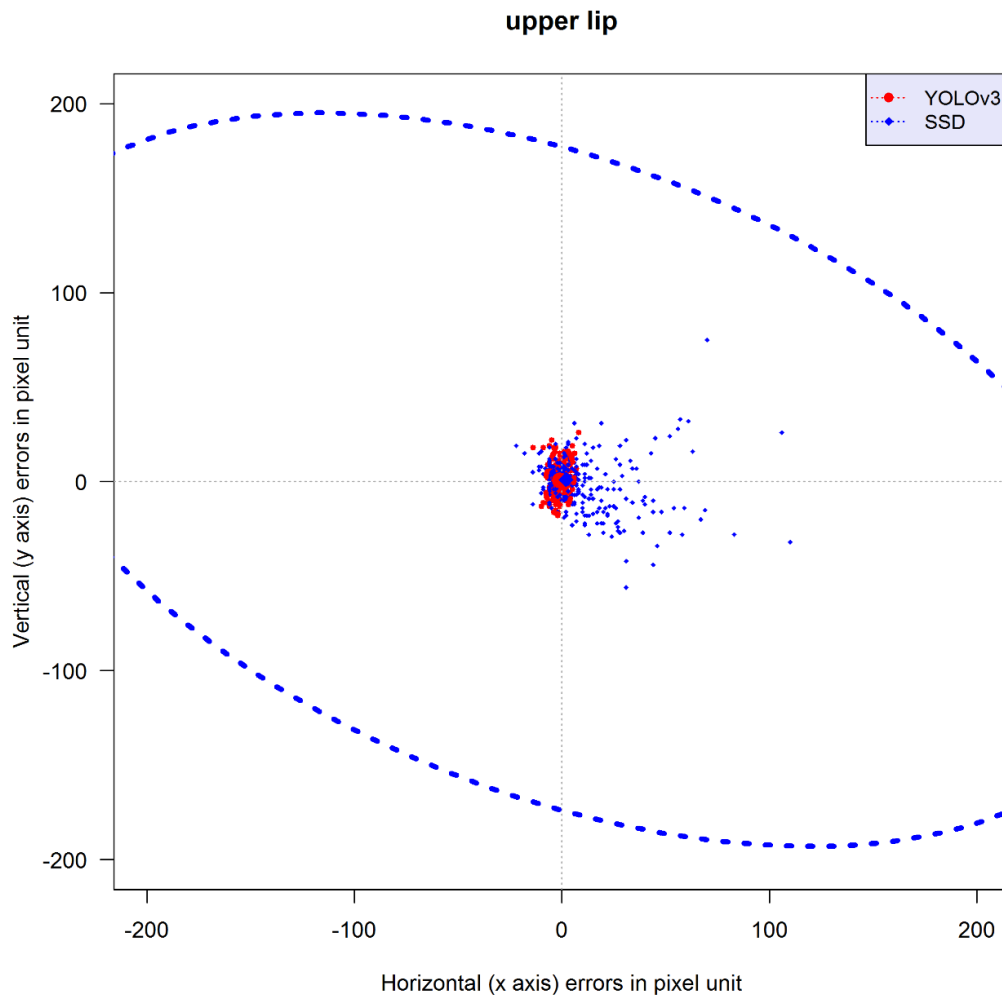
44

upper lip

Figure 5, R. Error scattergrams and 95% confidence ellipses for the error that were obtained from the YOLOv3 (red) and SSD (blue) methods of the point "upper lip". All the measurements were calculated by pixel. The plots clearly indicate that plots of YOLOv3 showed a smaller error range than SSD did.

국문초록

두부계측방사선 사진 계측점 자동 식별의 최신 기계 학습 알고리즘 간 정확도 및 연산 성능 비교 연구 – YOLOv3 vs SSD

박 지 훈

서울대학교 대학원 치의과학과 치과교정학 전공

(지도교수: 이 신 재)

연구 목적: 본 연구의 목적은 두부계측방사선 사진 계측점 자동 식별에 있어, 최근 개발된 두 가지 딥 러닝 알고리즘의 정확도와 연산 성능을 비교하는 것이다. 본 연구에서는 다음 두 가지의 알고리즘을 계측점 자동 식별에 적용하였다. 1) You-Only-Look-Once version 3 (YOLOv3) 및 2) the Single Shot Detector (SSD).

재료 및 방법: 총 1,028 개의 두부계측방사선 사진 영상이 YOLOv3 와 SSD 방식의 학습 데이터로 사용되었다. 대상 계측점은 80 개였다. 학습 과정을 거친 후, 각각의 알고리즘을 새로운 283 개의 테스트 영상에서 비교 분석하였다. 정확도는 1) 평균적인 point-to-point error, 2) success detection rate (SDR), 그리고 3) 2 차원 평면에서 시각화한 scattergram 을 기반으로 평가했다. 각각의 알고리즘의 평균 연산 시간 역시 기록하였다.

결과: YOLOv3 는 SSD 에 비해 총 38/80 개의 계측점에서 더 높은 정확도를 보였다. 나머지 42/80 개의 계측점은 두 알고리즘 간에 정확도에 있어

통계적으로 유의미한 차이를 나타내지 않았다. Error plot 에서는 YOLOv3 가 SSD 에 비해서 error 의 범위가 더 작을 뿐 아니라, 2 차원 평면에서 방향성의 영향을 덜 받는 것으로 나타났다. 하나의 영상에서 계측점을 자동 식별하는데 소요된 평균 시간은 YOLOv3 와 SSD 가 각각 0.05 초, 2.89 초로 기록되었다. 본 연구에서 YOLOv3 는 기존 문헌에서 최상의 정확도를 기록했던 연구에 비해 약 5% 가량 높은 정확도를 보였다.

결론: 본 연구를 통해 적용된 두 개의 알고리즘 중, YOLOv3 가 두부계측방사선 사진 계측점 완전 자동 식별의 임상적인 적용에 가능성 높은 알고리즘임을 확인하였다.