Tooth Localization with Coarse-to-Fine Auto-Encoders

Ayşe Betül OKTAY

Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Istanbul, Turkey
abetul.oktay@medeniyet.edu.tr
(Geliş/Received:31.05.2017; Kabul/Accepted:24.01.2018)
DOI: 10.17671/gazibtd.317893

***Abstract*—** Localization of teeth is a prerequisite task for most of the computerized methods for dental images such as medical diagnosis and human identification. Classical deep learning architectures like convolutional neural networks and auto-encoders seem to work well for tooth detection, however, it is non-trivial because of the large dental image size. In this study, a coarse-to-fine stacked auto-encoder architecture is presented for detection of teeth in dental panoramic images. The proposed architecture involves cascaded stacked auto-encoders where sizes of the input patches are increased with the successive steps. Only the detected candidate tooth patches are fed into the successive layers, thus the irrelevant patches are eliminated. The proposed architecture decreases the cost of detection process while providing precise localization. The method is tested and validated on a dataset containing 210 dental panoramic images and the tooth detection accuracy of the system is 91%.

***Keywords*—**Coarse-to-Fine auto-encoder, detection, deep learning, panoramic dental image

Büyükten Küçüğe Oto-Kodlayıcılar ile Dişlerin Konumlandırılması

***Özet*—** Dişlerin lokalizasyonu, bilgisayar destekli gerçekleştirilen diş görüntülerden insan kimliklendirme ve medikal tanı için bir ön şarttır. Konvolüsyonel sinir ağları ve oto-kodlayıcılar gibi klasik derin öğrenme mimarileri diş tanıma işlemi için başarılı gözükse de dental görüntülerin çok büyük olması nedeniyle tüm arama uzayının taranması mümkün gözükmemektedir. Bu çalışmada, büyükten-küçüğe yığınlanmış bir oto-kodlayıcı yapısı ile dental görüntülerden dişleri tanıyan bir sistem sunulmuştur. Önerilen mimari, girdi görüntü yamalarının boyutlarının her kademede arttığı bir kademeli yığınlanmış oto-kodlayıcı yapısı içerir. İlerdeki katmanlara sadece bulunan aday diş yamaları verilir; böylece alakasız yamalar elimine edilmiş olur. Önerilen mimari tanıma aşamasındaki maliyeti düşürmekle beraber hassas konumlandırma imkanı sunar. Geliştirilen metot, 210 dental panoramik görüntü içeren bir veri kümesi üzerinde test edilmiştir ve %91 doğruluk ile çalışmıştır.

***Anahtar Kelimeler*—**Büyükten –küçüğe oto-kodlayıcı, tanıma, derin öğrenme, panoramik dental görüntü

1. INTRODUCTION

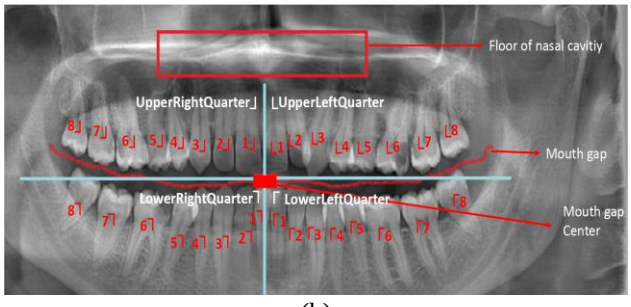
Detection of teeth from dental images [1] is the first step of most computerized dental applications like medical diagnosis [2] and human identification [3]. However, it is a challenging task because of the imaging artifacts and noise, large variation, and dental restorations. An example of a panoramic dental image is shown in Figure 1-a.

In the literature, most of the studies for tooth detection are performed with low-level vision techniques for intra-oral dental images that include a few teeth [4], [5]. There also exist machine learning based systems with hand-crafted features for tooth identification. For example, Mahoor and Abdel-Mottaleb [6] employed a Bayesian classifier to classify the teeth in bitewing images. Then, the labels of the teeth are determined using the spatial relationships between the teeth. The system presented in [7] implements the watershed algorithm to segment the

panoramic images into small regions and runs a fitness function with a set of features of each region for tooth detection. In the study of [3], support vector machines are used with several geometrical properties of the teeth for classification of the molar and the premolar teeth. Support vector machines are also employed for tooth classification and labeled them with the Markov Chain Model [8].

Recently, deep learning has been widely used in many computer vision applications and the state of the art results have been improved with the usage of training large amount of data [9], [10]. However, it is difficult to get such large training data (like 200M training images for face recognition [11]) for medical imaging applications. First of all, it is very hard to gather huge amount of real medical data from voluntary subjects with ethical approvals. In addition, it is very time-consuming to mark/delineate the images by an expert for obtaining the ground truth data. Also, the structures in medical images have high variability according to the illnesses, pathologies, etc. Despite these difficulties, there are many successful medical imaging applications based on deep learning in literature [12], [13] that produce successful results.

(a)



(b)

Figure 1. (a) An example of a dental panoramic image. (b) Cropped dental image where mouth gap, floor of nasal cavity and teeth are labeled.

Unsupervised learning of the tooth features seems to work well for detection with deep learning. Hierarchical tooth features can be learnt in an unsupervised manner with a deep architecture like stacked auto-encoders and teeth can be localized with a sliding window technique. However, the number of sliding windows would be very large during testing due to the big dental image sizes and rotational differences. For example, a dental panoramic image of 1500x2500 pixels size will have nearly 4M windows for each different window size and rotation which is intractable The dental image can be scaled down to smaller sizes as a solution but this process would be too coarse for precise localization.

In this paper, we propose a coarse-to-fine stacked auto-encoder framework for the detection of teeth in panoramic dental images. The proposed framework proposes an efficient way to detect the candidate tooth positions with cascading auto-encoders. The tooth candidates are coarsely determined at the first steps and the final localizations are performed at the final steps at finer resolutions. Thus, the running time decreases dramatically with the decreased searched space and detection is performed in a more controlled and tractable way.

Similar architectures that generate features at multiple resolutions are used for precise localization of objects. Honari et al. [14] stated that the max-pooling process discards spatial information for creating more robust features and proposed Recombinator Networks for facial keypoint localization. Their architecture employs a network that aggregates information from features across different levels of the network using concatenation. The studies [15], [16] proposed convolutional networks for localization and segmentation that generate features at different resolutions and combined these features to produce the final features. Zhang et al. [17] use coarse-to-fine auto-encoders for face alignment by using cascaded auto-encoders.

For panoramic dental images, it is intractable to detect all of the teeth simultaneously with a deep architecture because of the large image size. In addition, simultaneous detection of all teeth is difficult because of the missing teeth, dental restorations, and their repetitive structure. Therefore, we first find the possible tooth area in the images with a basic pre-processing steps. After reducing the search area, the teeth are detected with the cascaded stacked auto-encoders. The proposed system is tested on a dataset containing 200 images and the detection results are promising.

2. DENTAL PANORAMIC IMAGES AND PREPROCESSING

A panoramic dental image is an X-ray image (Figure 1-a) that includes the whole mouth. Normally, there are 8 teeth in each quarter q of mouth where the upper left, upper right, lower left, and lower right quarters are shown by the symbols $q \in \{\lrcorner\ \llcorner\ \urcorner\ \ulcorner\}$, respectively. We use Palmer's dental notation system where each tooth t_i, $1 \leq t_i \leq 8$, is numbered beginning from the mouth center with the quarter symbol q. For example, the upper right central incisor has the label $1\lrcorner$, while the upper left canine tooth has the label $\llcorner 3$. Figure 1-b shows the label of each tooth and the other structures floor of nasal cavity and mouth gap- in a panoramic dental X-ray image.

The panoramic images include structures like jaws, sinuses, etc. besides teeth. In order to reduce the size of the search space for detection of teeth, we first detect the possible placement of teeth according to the mouth gap. While a panoramic dental image is taken, the subject bites a plastic plate with front teeth to ensure that mouth is open and to fix the head position. Thus, the mouth gap center is close to the vertical center of the image. Consider I as an image of size mxn. Let $\phi(w^{i,j})$ be a function that gives the location of the minimum value of horizontal projection histogram of the window w centered at the image location $I(i,j)$. The mouth gap center $m_c(n/2)$ is found by

$$m_c\left(j=\frac{n}{2}\right)=\ \phi\left(w^{i,\frac{n}{2}}\right),\ j=n/2 \qquad (1)$$

where n is the length of the image I. In order to automatically determine i, we use the floor of nasal cavity f which is the brightest part of I on the upper front teeth. The maximum value of the projection histogram of the window centered at $i=m/4$ and $j=n/2$ is detected as the floor of nasal cavity.

The mouth gap m_c is detected for the $i > f + \tau$ where τ is a threshold. Detection of mouth gap center with respect to the f prevents the false detection of the sinuses (which are also dark pixels) as mouth gap. After finding the mouth gap center $m_c(n/2)$, the whole mouth gap is found iteratively to left and right by Eq. 2 and Eq. 3, respectively.

$$m_c(j+k)=argmin[\phi(w^{i,j+k})+\delta m_c(j+k-1)],\ if\ j>n/2 \qquad (2)$$

$$m_c(j-k)=argmin[\phi(w^{i,j-k})+\delta m_c(j+k+1)],\ if\ j<n/2 \qquad (3)$$

where k is the iteration step and δ is a weighting term. If the value of ϕ is greater than a threshold the algorithm stops. The final mouth gap is found by fitting a spline to the detected points. The dental images have different sizes and the images are resized according to the mouth gap length. This process reduces the need of using many sliding windows at different sizes. Then, the mouth quarters q {┘ └ ┐ ┌} each containing 8 teeth are automatically detected. Each quarter has different tooth appearance because of rotation, so other 3 quarters {┘ └ ┌ } are rotated and mirrored to make the similar appearance of teeth within the down left quarter ┐. Note that, this transformation provides training and testing all quarters together by eliminating the appearance differences between them.

3. STACKED-AUTO ENCODERS

Auto-encoders are neural networks that effectively learn the hierarchical representation of data. The stacked auto-encoders are stacked versions of auto-encoders and they are trained bottom up in unsupervised manner and features learned at the top layer are used for supervised training and fine-tuning.

Let $w^{i,j}$ be an image window and $x \in \mathbb{R}^d$ be the pixels in the window. An auto-encoder is a neural network that has an input, a hidden h, and an output layer. It takes x as input and tries to find the optimal parameters $\Theta = (W, b)$ by minimizing the amount of distortion between the input and output. The input x is mapped to a latent representation $h = (Wx + b)$ where the hidden layer h is a new feature representation of the input image x. Then, the output layer (decoder) is trained to reconstruct the input image patch from the hidden representation with reverse mapping training by back-propagation. After training, the weights W are used as the representation of data. The stacked auto-encoders are formed by stacking multiple auto-encoders by wiring output of each layer to the input of the successive layer.

4. COARSE-TO-FINE STACKED AUTO-ENCODERS

In the proposed framework, a number of stacked auto-encoders are cascaded to detect the teeth. The main objective of the architecture is making better detections at successive auto-encoders by the increasing image resolution and decreasing the running time. Only the candidates with better scores are evaluated at the successive layers which reduces the cost of detection. The architecture of the proposed framework is shown in Figure 2.

The first stacked auto-encoder takes the image patches of as input x and extracts the feature representation h. These randomly selected patches are down-scaled into small patch sizes where $kxk = x$. After the representations are learned, supervised learning is performed with a neural network layer with soft-max classifier that produces a score $s_{i,j}$ indicating the probability of being a tooth candidate. The candidates having the best scores are fed into the successive layers. The number of best candidates n^l with high scores at level l, that will be trained at the successive layer, are decreased while the size of the image patch is doubled. This allows to learn better tooth representations at successive layers and eliminates the non-tooth windows at the former stages.

5. TOOTH LOCALIZATION

In order to find the optimal final tooth positions T'=$\{t'_1, \dots, t'_8\}$, basic morphological operations are used. The n^3 candidates are the final candidates and maximum 8 teeth should be detected for each quarter. First, the detected pixels are grouped with labeling connected components and pixel groups are formed. A group g_k has the average probability score

$$S^{gk}=\frac{1}{n}\sum s_{i,j}, \qquad (4)$$

that is the sum of candidate scores $s_{i,j}$ divided by number of pixels in the group. Let the location of the centroid of

.

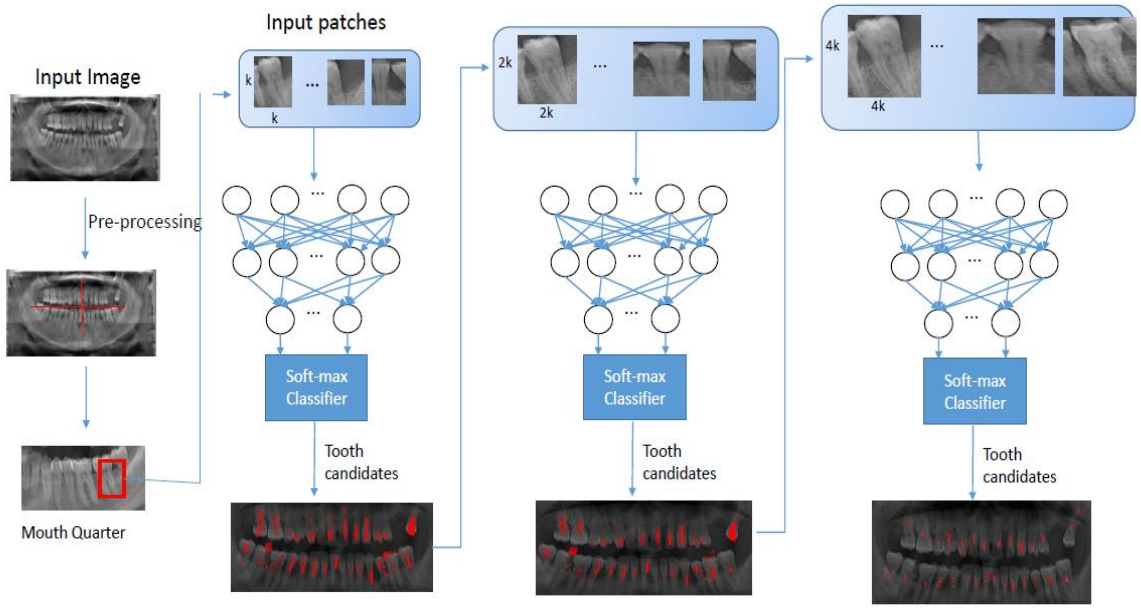


Figure 2. Framework of the proposed system

the group be $L(g_k)$. We model the tooth positions with a chain graph with 8 nodes. The final tooth positions are determined by

$$T' = \max_{S} \sum_{t_i=1,\dots,8} S^{gk} + \frac{|L(g_k)-\mu(t_i)|}{2\sigma_{t_i}^2} \qquad \textbf{(5)}$$

where μ is the mean and σ is the standard deviation of the Euclidean distance between tooth t_i and its neighboring tooth t_{i+1} which are calculated through the training set. The first term in Eq. 5 includes the tooth representation found by the stacked auto-encoders. The second term is the geometric term that is the difference between the group center and the average position of the tooth t_i according to the training set. The geometric term prevents two neighboring teeth being very close or far.

If the difference is greater than a threshold, the tooth is determined as a missing tooth. By solving Eq. 5 by dynamic programming, the optimal solution is found in polynomial time.

6. EXPERIMENTS

In order to evaluate the effectiveness of the proposed framework, a dataset including 210 dental panoramic images of 174 different subjects is used. The images are taken by 3 different dental panoramic X-ray machines and they have different image sizes which are 2871x1577, 1435x791, and 2612x1244. The ground-truth bounding boxes for each tooth are delineated by an expert. There are 5568 teeth and number of missing teeth is 829. There are abnormalities like impacted tooth and many dental works like crowns, fillings, implants, and braces in the images.

In the pre-processing step, the window size w for detection of the mouth gap is selected as 800x100 pixels. The weighting term λ is 0.2 and the threshold t is 150 pixels. After the mouth gap is detected, all of the images are scaled to the maximum mouth gap length by preserving the aspect ratio.

For training the system, 10 of images of 10 different subjects are used. 50000 patches that have image sizes between the minimum and maximum tooth sizes are randomly selected. All of the patches are down-scaled to 32 by 32 pixels for the first level of the auto-encoders where k = 32. For the supervised soft-max classifier, 150 windows that contain a whole tooth are used as positive samples and 300 images that contain at most 30% of a tooth are used as negative samples. The number of candidates n^1 is 3000 windows for the first level, n^2 = 1500 for the second level and n^3 = 500 for the third level.

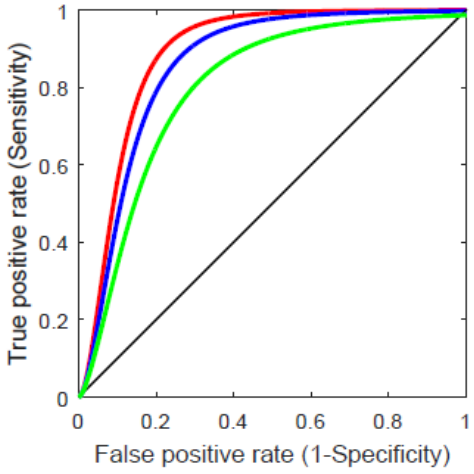


Figure 3. ROC of each auto-encoder level.

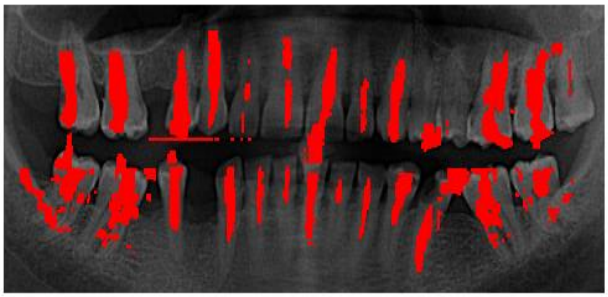
Table 1. Detection percentages for each tooth

Tooth number	1	2	3	4	5	6	7	8
⌞	*88*	*93*	*91*	*89*	*92*	*92*	*91*	*89*
⌜	*89*	*89*	*87*	*90*	*91*	*93*	*91*	*87*
⌟	*91*	*92*	*89*	*91*	*89*	*90*	*93*	*89*
⌝	*90*	*93*	*86*	*92*	*91*	*91*	*88*	*84*

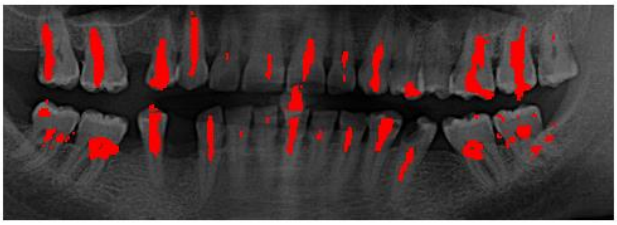
In order to evaluate the effectiveness of our system, we first detect the accuracy of the each level at the coarse-to-ne framework. A detection is evaluated as accurate if the window includes at least 50% of the box including the tooth delineated by the expert. The ROC curves of each level is shown in Figure 3. The accuracy rate increases at each level while the true positive rate is 0.81, 0.89, and 0.91, at the levels 1, 2 and 3. The increasing true positive rates show the effectiveness of our system.

In order to evaluate the final localization results of the post-processing step, we calculated the average detection accuracy. If the finally detected tooth t_0 is fully inside the ground truth, it is evaluated as an accurate detection. The detection percentages for each tooth is shown in Table 1. The down first teeth have the lowest detection rates. This may cause because of the brightness of the plastic plate occluding the front teeth. Molar teeth numbered 6, 7, and 8 have the highest percentages. The visual results are shown in Figure 4. Figure 4-a shows the tested panoramic dental image. Figure 4-b and Figure 4-b show the detection results of the first and second level, respectively. The detected pixels detected as teeth are refined in Figure 4-c and final localization results are shown in Figure 4-d.

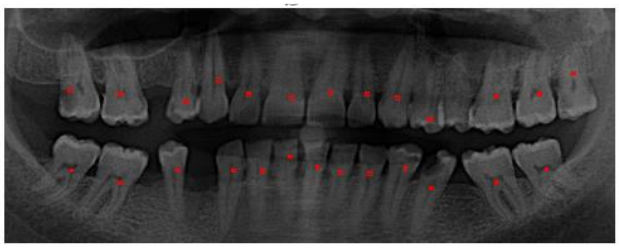
(a)



(b)



(c)



(d)

Fig. 4: a) A dental panoramic image with missing teeth. b) shows the detection results of the first level and c) shows the detection results of the last level of coarse-to-fine auto-encoders. d) Final detection results. Note that, the left upper fifth tooth couldn't be detected.

7. CONCLUSIONS

In this paper, a coarse-to-fine stacked auto-encoder is presented for detection of teeth in dental panoramic images. The presented technique has three important advantages. First, the search space is reduced with the developed mouth gap based pre-processing technique. Also, transforming the mouth quarters into similar appearance provided more samples for training while eliminating the cost of training each quarter separately. Second, the cascaded auto-encoders with increasing input image sizes produced precise localization results while decreasing the search time. Third, the patches not involving a tooth are effectively eliminated at the first steps. In the future, we will use the detected teeth and their hierarchical features for human identification by comparing same subjects' dental images.

REFERENCES

[1] P.-L. Lin, P.-W. Huang, Y. S. Cho, and C.-H. Kuo. An automatic and effective tooth isolation method for dental radiographs. Opto-Electronics Review, 21:126-136, 2013.

[2] A. Katsumata and H. Fujita. Progress of computer-aided detection/diagnosis (cad) in dentistry. Japanese Dental Science Review, 50(3):63 - 68, 2014

[3] P. Lin, Y. Lai, and P. Huang. An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information. Pattern Recognition, 43(4):1380 - 1392, 2010.

[4] J. Zhou and M. Abdel-Mottaleb. A content-based system for human identification based on bitewing dental x-ray images. Pattern Recognition, 38(11), 2005.

[5] M. Abdel-Mottaleb, O. Nomir, D. Nassar, G. Fahmy, and H. Ammar. Challenges of developing an automated dental identification system. In Circuits and Systems, 2003 IEEE 46th Midwest Symposium on, volume 1, pages 411-414, 2003.

[6] M. H. Mahoor and M. Abdel-Mottaleb. Classification and numbering of teeth in dental bitewing images. Pattern Recognition, 38(4):577 - 586, 2005

[7] D. Frejlichowski and R. Wanat. Extraction of teeth shapes from orthopantomograms for forensic human identi_cation. In P. Real, D. Diaz-Pernil, H. Molina Abril, A. Berciano, and W. Kropatsch, editors, Computer Analysis of Images and Patterns, volume 6855 of Lecture Notes in Computer Science, pages 65-72. Springer Berlin Heidelberg, 2011.

[8] A. K. Jain and H. Chen. Registration of dental atlas to radiographs for human identification. In A. K. Jain and N. K. Ratha, editors, Biometric Technology for Human Identification II, volume 5779 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, pages 292-298, Mar. 2005.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 1701-1708, June 2014.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211-252, 2015.

[11] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[12] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer istopathology images. IEEE Transactions on Medical Imaging, 35(1):119-130, Jan 2016.

[13] H.-C. Shin, M. Orton, D. Collins, S. Doran, and M. Leach. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(8):1930-1943, 2013.

[14] S. Honari, J. Yosinski, P. Vincent, and C. J. Pal. Recombinator networks: Learning coarse-to-_ne feature aggregation. June 2015.

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[16] B. Hariharan, P. A. Arbel_aez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and _ne-grained localization. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pages 447-456, 2015.

[17] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In Computer Vision - ECCV 2014 - 13th European Conference, Proceedings, Part II, pages 1-16, 2014

.