Mask R-CNN ile Derin Yüz Sezici Gerçekleme
Design of a Deep Face Detector by Mask R-CNN

Ozan Cakiroglu Caner Ozer Bilge Gunsel
Çoğulortam Sinyal İşleme ve Örüntü Tanıma Grubu
Elektronik ve Hab. Müh. Bölümü, İstanbul Teknik Üniversitesi, Türkiye
{cakirogluoz, ozerc, gunselb}@itu.edu.tr

***Özetçe* —Bu çalışmada literatürde varolan Mask R-CNN derin öğrenme ağı yüz sezme amaçlı eğitilmiş ve öğrenilen yüz nesnesi için başarım raporlanmıştır. Çalışmada, varolan çalışmalardan farklı olarak, az sayıda yüz gözlemi ile bir eğitim gerçeklenmiş ve yüz nesnesini çevreleyen kutunun yakalanmasının yanı sıra, yüz bölgesinin piksel bazında bölütlenmesi de hedeflenmiştir. Eğitim PASCAL-VOC veri setinden alınan 2695 adet gözlem ile gerçeklenmiştir. Başarım, yüz sezme çalışmalarında referans alınan WIDER FACE veri tabanı üzerinde 159.000 yüz nesnesi için raporlanmıştır. Sonuçlar, baz alınan çalışmaya [1] göre çok küçük ölçekli yüzleri sezmede %6, orta ölçekli yüz sezmede %12, büyük ölçekli yüz sezmede %3 daha yüksek başarım sağladığını göstermektedir. Performansımızın Viola & Jones yüz seziciden daha yüksek olduğu da raporlanmaktadır. Bu çalışmada derin yüz sezici eğitiminde kullanılan öğrenme örnekleri için çıkarılan bölütleme verisi ve TensorFlow ortamında gerçeklenen eğtim-test rutinleri GitHub altında genel kullanıma açılmıştır.**

Anahtar Kelimeler*—*Yüz sezme, örnek düzeyinde bölütleme, derin öğrenme, evrişimsel sinir ağları.

***Abstract*—In this work an existing object detector, Mask R-CNN, is trained for face detection and performance results are reported by using the learned model. Differing from the existing work, it is aimed to train the deep detector with a small number of training examples and also to perform instance segmentation along with an object bounding box detection. Training set includes 2695 face examples collected from PASCAL-VOC database. Performance has been reported on 159,000 test faces of WIDER FACE benchmarking database. Numerical results demonstrate that the trained Mask R-CNN provides higher detection rates with respect to the baseline detector [1], particularly 6%, 12%, and 3% higher face detection accuracy for the small, medium and large scale faces, respectively. It is also reported that our performance outperforms Viola & Jones face detector. We released the face segmentation ground-truth data that was used to train Mask R-CNN and training-test routines developed in TensorFlow platform to public usage at our GitHub repository.**

Keywords*—*Face detection, instance segmentation, deep learning, convolutional neural networks.

I. Giriş

Son yıllarda, yüz sezme sistemlerinde evrişimsel sinir ağ temelli tek [2], [3] veya iki [4], [5] aşamalı nesne sezme yöntemleri kullanılmaktadır. Bu yöntemlerde farklı ölçeklerdeki yüzlerin bulunması amaçlanmış ve klasik yüz sezicilere [6] göre daha üstün başarım oranı elde edilmiştir. Ayrıca, RetinaNet [7] veya çevrimiçi keskin örnek madenciliği (OHEM) [8] kullanılarak, eğitimde karşılaşılan keskin (hard) örneklerin daha iyi öğrenilmesi sağlanabilmektedir. Genel olarak öğrenme aşamasında yüz nesnesinin derin yüz sezici tarafından öğrenilmesi sağlanmakta, ardından başarım farklı veri tabanlarında raporlanmaktadır. Bu çalışmalarda, yüz sezme başarımının raporlanması amacıyla WIDER FACE [1], FDDB [9] ve PASCAL-VOC Faces [10] veri tabanları sıklıkla kullanılmaktadır. Başarım, ortalama keskinlik (AP) cinsinden raporlanmakta ve başarımın büyük oranda yüz nesnesinin ölçeği arttıkça yükseldiği çıkarımı raporlanmaktadır.

Bu çalışmada, literatürde varolan yüz sezicilerden farklı olarak, örnek düzeyinde bölütleme de yaparak yüz sezme başarımını arttırmak amacıyla Mask R-CNN derin öğrenme ağı eğitilmiş, öğrenme ve test başarımı literatüre uygun olarak raporlanmıştır.

II. Mask R-CNN ile Nesne Sezme

Bu bölümde Mask R-CNN [11] derin sinir ağı ile giriş görüntüsü üzerinde gerçeklenen iki aşamalı nesne sezme ve öğrenme modelinin çıkarılımı anlatılmaktadır. Mask R-CNN mimarisinde ilk aşamada olası nesne bölgelerinin koordinatları ve nesne içerme olasılıkları, bölge öneri ağı (RPN) adı verilen yapı ile çıkartılır. İkinci aşamada her bir olası nesne bölgesinin yüz içerme olasılığı ve nesneyi çevreleyen kutular (bounding box-BB) bulunur. Mask R-CNN örnek bazında bölütleme gerçekleştirerek BB lerin yanısıra yüz olarak etiketlenmiş piksellerden oluşan maskeyi de üretir. Şekil 1'de Mask R-CNN'in genel mimarisi üzerinde işlem akışı gösterilmektedir.

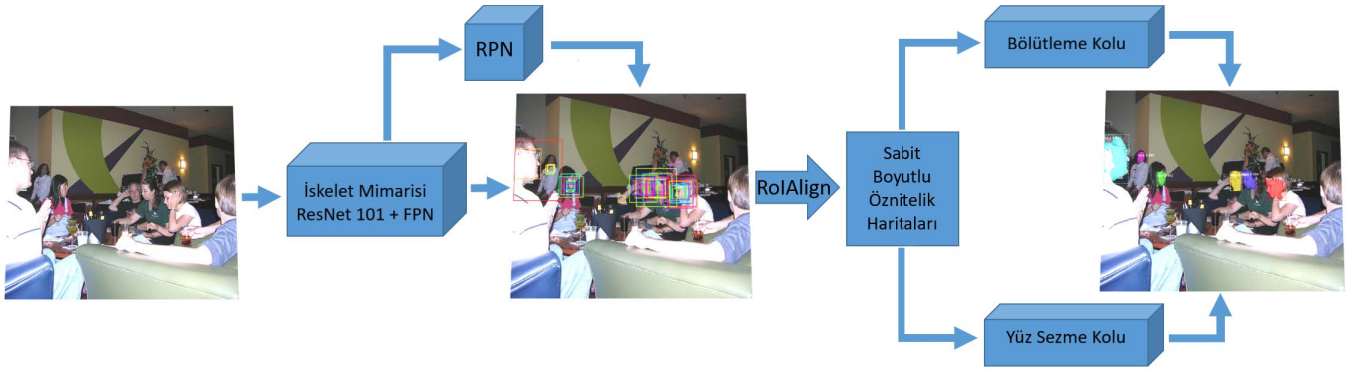
Verilen bir görüntü için konvolüsyonel sinir ağı iskelet mimarisi, öznitelik çıkarma işlemini gerçekleştirmektedir. Bu çalışmada öznitelik çıkarıcı olarak, farklı uzamsal boyutlarda öznitelik çıkartmakta sıklıkla kullanılan, ResNet 101'in [12] farklı seviyelerdeki çıktılarını kullanan Öznitelik Piramit Ağları (FPN) [13] yapısı kullanılmıştır. Elde edilen öznitelik haritaları RPN'den geçirilerek, bütün lokal bölgelerin nesne olabilirlik skorları ve BB koordinatları, sırasıyla ikili sınıflandırma ve regresyon ile bulunmaktadır.

Eğitim esnasında Eşitlik 1'de görülen RPN kayıp fonksiyonunun enküçüklenmesi sağlanır. Burada ilk terim sınıflandırıcı hatasını, ikinci terim ise regresyon hatasını modellemektedir.

$$L_{RPN} = \frac{1}{N}\sum_{i=1}^{N}(L_{cls}(p_i, p_i^*) + L_{reg}(\boldsymbol{t_i}, \boldsymbol{t_i^*})). \qquad (1)$$

Eşitlik 1'de N eğitim sırasında kullanılan ankor sayısını, p_i ve p_i^* sırasıyla, yakalanan BB içerisinde nesne olabilirlik skorunun tahmin edilen ve gerçek referans değerini göstermektedir. Referans ile örtüşme oranına göre pozitif için 1,

978-1-7281-1904-5/19/$31.00 ~©~2019 IEEE



Şekil 1: Mask R-CNN'in genel yapısı.

negatif için 0 olmak üzere, L_{cls} sınıflandırıcının ikili çapraz entropi kayıp fonksiyonunu modeller. $\boldsymbol{t_i}$ ve $\boldsymbol{t_i^*}$, sırasıyla, nesne BB sine ait tahmin edilen ve gerçek koordinat vektörleridir. L_{reg} regresyon ile belirlenen dikdörtgensel BB ye karşı düşen Smooth L_1 [14] kayıp fonksiyonunu göstermektedir. Eğitim esnasında L_{RPN}'nin enküçüklenmesi sağlanarak daha iyi öznitelik çıkarımı ve nerelerde nesne olduğunun daha iyi öğrenilmesi sağlanır.

İkinci aşamada, en yüksek nesne bulunma olasılığına sahip olan bölgelerin uygun öznitelik haritası üzerinde denk düştüğü alanlar, RoIAlign ile sabit bir boyuta getirilir. Sabit boyutlu öznitelik parçaları kullanılarak iki farklı kol vasıtasıyla, görüntü içerisinde sezilen BB'lerin yüz içerme olasılığı, görüntü içerisindeki lokasyonları ve Mask R-CNN'e has bölütleme haritaları çıkartılır. Yüz veya yüz değil şeklinde ikili sınıflandırma ve kesinleştirilmiş lokalizasyon işlemleri için 7x7'lik uzamsal boyuttaki öznitelik haritaları kullanılır. Bölütleme haritalarını çıkarmak için ise tümden-evrişimsel bir ağ ile 256 kanallı 14x14'lük uzamsal boyuttaki öznitelik haritaları kullanılmaktadır.

RPN tarafından önerilen nesne olma olasılıklı BB'lere karşı düşen öznitelik parçaları, yakalanan nesne sınıfının ve bölütleme maskesinin belirlenmesi amacıyla, Mask R-CNN mimarisinde bulunan üst ağ içerisinde ikinci bir en iyilemeye sokulur. Bu en iyilemede kullanılan kayıp fonksiyonu Eşitlik 2'de verilmektedir. Mask R-CNN'in bölütleme, regresyon ve ikili sınıflandırıcı kayıp fonksiyonu L_{Head} içerisinde tümleşik halde tanımlanmıştır. Bunun nedeni Eşitlik L_{RPN}'in en küçüklenmesi ile elde edilen BB bilgilerinin Eşitlik 2'de kullanılıyor olmasıdır.

$$L_{Head} = \frac{1}{M}\sum_{i=1}^{M}(L_{cls}(y_i, y_i^*) + L_{reg}(\boldsymbol{t_i}, \boldsymbol{t_i^*}) + L_{mask}(T_i, T_i^*)) \tag{2}$$

Eşitlik 2'de eğitim için kullanılan ve RPN ile oluşturulan nesne içeren bölge sayısı M, tahmin edilen bölgenin yüz olarak sınıflandırılma olasılığı için tahmin edilen ve gerçek değerler sırasıyla, y_i ve y_i^*, tahmin edilen ve gerçek nesne BB'lerine ilişkinkoordinat vektörleri sırasıyla, t_i ve t_i^*, tahmin edilen ve gerçek referans bölütlemeleri sırasıyla, T_i ve T_i^* ile gösterilmiştir. L_{cls} ve L_{reg} eşitlik 1'de tanımlı fonksiyonlardır. L_{mask} Eşitlik 3'te tanımlanmıştır.

$$\begin{aligned} L_{mask}(T_i, T_i^*) = -\frac{1}{H}\frac{1}{W}\sum_{h=1}^{H}\sum_{w=1}^{W} & T_{i,(h,w)}^* log(T_{i,(h,w)}) \\ & -(1 - T_{i,(h,w)}^*) log(1 - T_{i,(h,w)}) \end{aligned} \tag{3}$$

Eşitlik 3'te H ve W bölütleme haritasının yüksekliğini ve genişliğini, $T_{i,(h,w)}^*$ ve $T_{i,(h,w)}$, sırasıyla, i. bölge için yapılan bölütleme tahminini ve gerçek referans bölütlemeyi gösterir. Burada bölütleme haritası üzerinde her bir piksel için yapılan tahminin, ikili çapraz entropi kayıp fonksiyonu ile maliyeti bulunur.

Mask R-CNN'in uçtan uca eğitilebilmesi için Eşitlik 1 ve 2 toplanır ve ağırlıklara göre gradyanlar hesaplanır. Bu iki kayıp teriminin toplamının L ile gösterilmesi durumunda, $\nabla_W L$ kayıp fonksiyonunun ağırlıklara göre gradyanını gösterecektir. Eşitlik 4 ağırlıkların güncellenmesi için kullanılan momentum yöntemini göstermektedir.

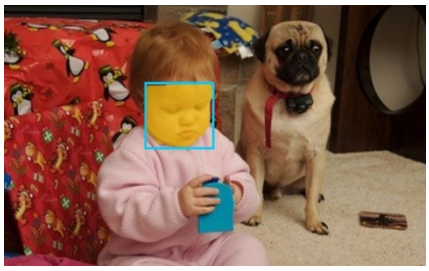
$$\begin{aligned} x &:= \mu x - \eta \nabla_W L \\ W &:= W + x \end{aligned} \tag{4}$$

Eşitlik 4'de x gradyanın değişiminin hızını, μ momentum katsayısını, η ise öğrenme katsayısını göstermektedir.

III. Yüz Sezici Eğitimi

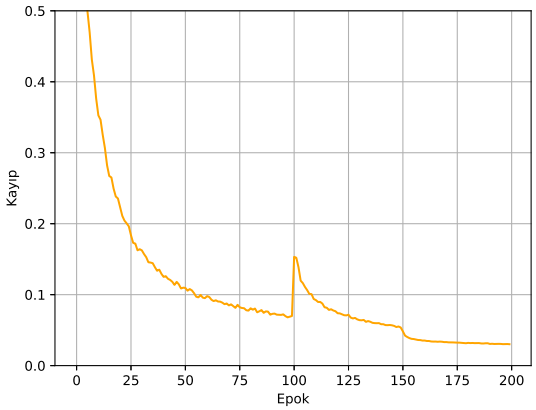
Çalışmamızda Bölüm II'de mimarisi verilen Mask R-CNN derin yüz sezici olarak eğitilmiştir. Hedeflenen, 80 farklı nesneyi sezmek üzere eğitilmiş olan Mask R-CNN nesne sezicisinin, yeni bir nesne olarak tanımlanan yüzleri sezebilme amacıyla, küçük bir eğitim seti ile eğitilmesidir. Bu doğrultuda öncelikle PASCAL-VOC veri setinden [10] 2695 yüz örneği barındıran 1454 görüntü ile eğitim seti oluşturulmuştur. Mask R-CNN nesneyi çevreleyen dörtgen kutu (BB) sezimi yanı sıra, nesne bazında bölütleme de gerçekleştiğinden eğitim setindeki her örnek için BB gerçek referansı (BB-GT) ve bölütleme gerçek referansı (S-GT) gerekmektedir. Bu amaca uygun bir eğitim veri seti literatürde bulunmadığından çalışmamızda öncelikle BB-GT ve S-GT verileri oluşturulmuştur. Yüz sınıfına ait S-GT işaretleme verisinin elde edilmesi için yüz nesnelerinin her bir örnek için piksel bazında bölütlenmesi gerçeklenmiştir. Bu işlem sahne öğrenme tabanlı LabelMe Annotation Tool [15] ile yarı-otomatik bir şekilde yapılmış

böylelikle başka grupların çalışmalarında da kullanılabilecek açık bir eğitim veri seti elde edilmiştir. Şekil 2'de oluşturulan veri setinden bir yüz örneği, kendisini çevreleyen BB ve piksel bazında işaretli bölütleme maskesi ile birlikte görülmektedir.



Şekil 2: Veri seti içerisinden işaretlenmiş görüntü örneği.

Eğitimler, NVIDIA GTX 1080 ekran kartıyla, 200 epok (epoch) ve her epokta 1400 görüntü örnekleyen iterasyonlar ile toplam 36 saatte tamamlanmıştır. Referans model olarak, Matterport'un [16] TensorFlow ve Keras kütüphaneleri ile yazılım ortamında gerçeklediği derinlikli öğrenme ağ yapısı alınarak, bu yapıya gerekli eklemeler yapılmıştır. Eğitim, ağ yapısının baş katmanlarının (head layers) 100 epok boyunca eğitilmesi ve ResNet 101 iskelet katmanları ve alt katmanların 50 epok eğitilmesi ile gerçeklenmiştir. Son olarak ağ yapısı uçtan uca 50 epok boyunca eğitilerek, öğrenme modeli elde edilmiştir. Mimarinin baş katmanlarının eğitiminde öğrenme oranı 0.001 olarak seçilirken, 100 epok ardından 0.0001'e düşürülmüştür. ResNet-101 iskelet mimarisinin ağırlıkları, MS COCO [17] veri setinde eğitilen ağırlıklar olarak belirlenerek, eğitimde başlangıç ağırlık değerleri olarak alınmıştır. Ankor boyutları (0.5, 1, 2) en-boy oranı ve (32, 64, 128, 256, 512)'lık pencere boyutları olmak üzere 5 adet olarak seçilmiştir. Tanımlanan yüz sezici eğitimi sırasında kayıp fonksiyonunun epok sayısına bağlı değişimi kaydedilmiştir ve Şekil 3'de görülmektedir. Grafikten görüldüğü üzere, kayıp azalarak minimum değerine yakınsamaktadır ki bu durum Mask R-CNN'in yüz nesnesini öğrendiğini göstermektedir.



Şekil 3: En iyileme kayıp fonksiyonun yapılan eğitim sırasında raporlanan değişimi.

IV. Başarım Test Sonuçları

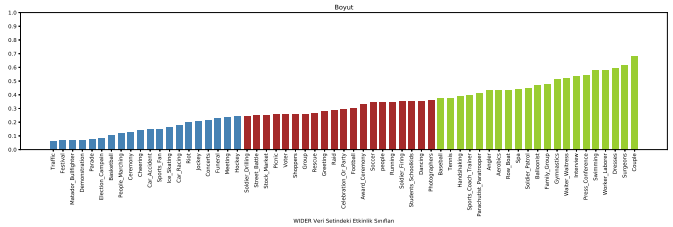
Testlerde öncelikle yüz sezme gerçeklemek üzere eğitilen derin sinir ağının öğrenme başarımı irdelenmiştir. Bu amaçla Pascal VOC üzerinde eğitim sonucu elde edilen model kullanılarak ve eğitim seti test seti olarak alınarak sezme başarımı raporlanmıştır. Tablo I'de yüz sezme başarımı duyarlılık (R-recall), kesinlik (P-precision) ve F-1 ölçütü ile farklı kesişim (IoU) eşik değerleri için raporlanmaktadır. Literatürde IoU=0.5 objenin doğru yakalandığı kabul edilen en düşük kesişim değeridir. Tablo I'de Tr-Tr altında raporlanan değerlerden görüldüğü üzere, F-1 ölçütüne göre başarım IoU $\geq$ 0.5 için %92 gibi yüksek bir değerdir ve bu durum eğitim sonucunda yüz objesinin başarıyla öğrenildiğini göstermektedir. Başarım yüksek IoU değerlerine gidildikçe azalmaktadır; ancak sezilen yüz sınırlarının bilinen gerçek referans ile en az %80 oranında örtüşmesi anlamına gelen IoU $\geq$ 0.8 değeri için halen %74 oranında düşük olmayan bir sezme başarımı raporlanmaktadır. İkinci test senaryosunda, Pascal VOC'da bulunan ancak eğitim seti ile örtüşmeyen 160 görüntüde bulunan 225 yüz üzerinde başarım raporlanmıştır. Tablo I'de Tr-Test altında farklı IoU eşik değerleri için raporlanan sonuçlar incelendiğinde, eğitilen yüz sezicinin, eğitimde görülmeyen görüntülerdeki yüz objelerini de hemen hemen aynı doğrulukla yakalayabildiği görülmektedir. Bu nedenle, modelin eğitim setini ezberlemediği ve öğrenilen ağ ağırlıklarının derin yüz seziciyi genellenebilir bir sezme performansına getirdiği sonucu çıkarılabilir.

		Tr-Tr			Tr-Test	
IoU	**R**	**P**	**F-1**	**R**	**P**	**F-1**
0.5	0.93	0.90	0.921	0.93	0.92	0.93
0.6	0.91	0.88	0.897	0.92	0.91	0.92
0.7	0.84	0.81	0.827	0.83	0.82	0.82
0.8	0.76	0.73	0.749	0.72	0.71	0.71

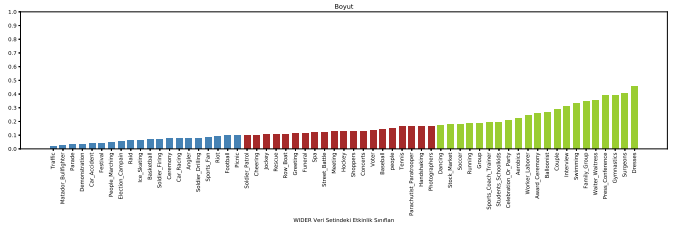
TABLO I: VOC Eğitim seti ve test seti üzerinde raporlanan yüz sezme başarımı.

Eğitilen yüz sezicinin başarımı, yüz sezmede oldukça zorlu bir veri seti olan WIDER FACE veri seti [1] üzerinde de sınanmıştır. WIDER veri seti 60 etkinlik sınıfından toplam 12.500 görüntü üzerinde 159.000 yüz içermektedir ve çok küçük ölçek, örtüşme, başın dönüklüğü gibi yüz sezmeyi zorlaştırıcı özellikleri nedeniyle başarım testlerinde referans alınan veri setlerindendir. Şekil 4'te eğitilen derin yüz sezici modelinin her bir etkinlik sınıfı klasörü için sağladığı performansı sezme oranı cinsinden gösteren histogram raporlanmaktadır. Sezme oranı, etkinlik sınıfında doğru sezilen yüz sayısının etkinlik sınıfında bulunan tüm yüzlerin sayısına oranı şeklinde hesaplanmaktadır. Zorluk derecelerini belirlemek açısından histogramda yatay eksendeki sınıflar içerdikleri yüz nesnelerinin ölçeklerine göre sıralanmıştır. Küçük ölçekli yüz nesnelerinin çoğunlukta bulunduğu etkinlik sınıfları mavi ile gösterilirken, kırmızı ve yeşil sırasıyla orta ve büyük ölçekli yüz nesnelerini çoğunlukla içeren sınıfları göstermektedir. Şekil 4'te görüldüğü gibi, modelimizin yüz sezme başarımı yüz objesinin ölçeği büyüdükçe artmaktadır. Karşılaştırma açısından Şekil 5'te Viola & Jones [6] yüz sezici ile elde edilen başarım, 6'da WIDER veri seti üzerinde elde edilen başarım için literatürde raporlanan sonuçlar paylaşılmaktadır. Tüm yöntemler için yüz nesnesinin ölçeği büyüdükçe başarımın yükseldiği görülmektedir. Ancak, Şekil 4 ve 6 karşılaştırıldığında, [1]'de

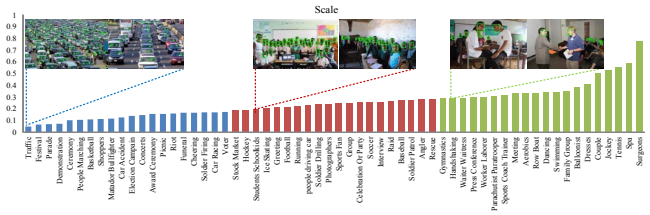
raporlanan yüz sezme başarımının, zor, orta-zor ve kolay olarak nitelenen üç farklı yüz ölçeğindeki sınıflar için sırasıyla % 5-19, %20-27, ve %27-75 aralıklarında kalmasına karşın, çalışmada eğitilen Mask R-CNN yüz sezici ile karşı düşen başarımların sırasıyla %5-26, % 26-40 , %40-78 aralıklarına yükseltilebildiği görülmektedir. Şekil 4 ve 5 karşılaştırıldığında ise eğittiğimiz modelin Viola & Jones yüz sezicisine göre başarımı, küçük ölçekli yüzler için %3 ile %11, orta ölçekli yüzler için %7 ile %28, büyük ölçekli yüzler için ise %12 ile %39 oranında arttırdığı görülmektedir.



Şekil 4: Eğitilen Mask R-CNN ile farklı ölçeklerde yüz sezme başarımının etkinlik sınıflarına göre değişimi.



Şekil 5: Viola & Jones [6] yüz sezicisinin yüz sezme başarımı



Şekil 6: [1]'deki yüz sezici ile raporlanan etkinlik sınıfları bazında sezme oranları.

V. Sonuçlar

Bu çalışmada Mask R-CNN sınırlı sayıdaki yüz gözlemi ile eğitilerek bir derin yüz sezici gerçeklenmiştir. Yüz sezme başarımı Pascal VOC ve WIDER FACE veri tabanları üzerinde raporlanmıştır. Literatürde COCO için 80 farklı nesne sezmek amacıyla eğitilmiş olan Mask R-CNN'in yüz nesnesini de sezebilmesine olanak sağlanmıştır. Kullanılan eğtim veri setine ilişkin gerçek referans bilgisi ve yazılım rutinleri kullanıma açılmıştır. Çalışmalarımız derin nesne seziciye daha farklı nesnelerin öğretilmesi ve sezme performansının arttırılması doğrultusunda devam etmektedir.

Teşekkür: Bu çalışmada Mask R-CNN eğitimine katkısı geçen Müh. Can Elbirlik'e teşekkür ediyoruz.

Kaynakça

[1] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *IEEE CVPR, pp.5525–5533*, 2016.

[2] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: single stage headless face detector," in *IEEE ICCV, pp. 4885–4894*, 2017.

[3] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3fd: Single shot scale-invariant face detector," in *IEEE ICCV, pp. 192–201*, 2017.

[4] H. Jiang and E. G. Learned-Miller, "Face detection with the faster R-CNN," in *IEEE FG, pp. 650–657*, 2017.

[5] S. Zhang, R. Zhu, X. Wang, H. Shi, T. Fu, S. Wang, T. Mei, and S. Z. Li, "Improved selective refinement network for face detection," *CoRR*, vol. abs/1901.06651, 2019.

[6] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[7] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE ICCV, pp. 2999–3007*, 2017.

[8] A. Shrivastava, A. Gupta, and R. B. Girshick, "Training region-based object detectors with online hard example mining," in *IEEE CVPR, pp. 761–769*, 2016.

[9] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[11] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE ICCV, pp. 2980–2988*, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR, pp. 770–778*, 2016.

[13] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *IEEE CVPR, pp. 936–944*, 2017.

[14] R. B. Girshick, "Fast R-CNN," in *IEEE ICCV, pp. 1440–1448*, 2015.

[15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. Journal of Computer Vision*, vol. 77, no. 1, pp. 157–173, May 2008.

[16] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," https://github.com/matterport/Mask_RCNN, 2017.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.