
**T.C.**
**GEBZE TEKNİK ÜNİVERSİTESİ**

**Bilgisayar Mühendisliği Bölümü**

# İKİ BOYUTLU GÖRÜNTÜLERDEN POZ TAHMİNİ

**Ali Yasin ESER**

**Danışman**
**Prof. Dr. Yusuf Sinan AKGÜL**

**Ocak, 2018**
**Gebze, KOCAELİ**


**T.C.**
**GEBZE TEKNİK ÜNİVERSİTESİ**

**Bilgisayar Mühendisliği Bölümü**

# İKİ BOYUTLU GÖRÜNTÜLERDEN POZ TAHMİNİ

**Ali Yasin ESER**

**Danışman**
**Prof. Dr. Yusuf Sinan AKGÜL**

**Ocak, 2018**
**Gebze, KOCAELİ**

Bu çalışma 4/01/2018 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Bölümünde Lisans Bitirme Projesi olarak kabul edilmiştir.

Bitirme Projesi Jürisi

| | | |
|---|---|---|
| Danışman Adı | Yusuf Sinan Akgül | |
| Üniversite | Gebze Teknik Üniversitesi | |
| Fakülte | Mühendislik Fakültesi | |

| | | |
|---|---|---|
| Jüri Adı | Erchan Aptoula | |
| Üniversite | Gebze Teknik Üniversitesi | |
| Fakülte | Mühendislik Fakültesi | |

# ÖNSÖZ

Bu raporun ilk taslaklarının hazırlanmasında emeği geçenlere, raporun son halini almasında yol gösterici olan Sayın Prof. Dr. Yusuf Sinan Akgül hocama ve bu çalışmayı destekleyen Gebze Teknik Üniversitesi'ne içten teşekkürlerimi sunarım.

Ayrıca eğitimim süresince bana her konuda tam destek veren aileme ve bana hayatlarıyla örnek olan tüm hocalarıma saygı ve sevgilerimi sunarım.

**Ocak, 2018** **Ali Yasin ESER**

## İÇİNDEKİLER

# ŞEKİL LİSTESİ

# TABLO LİSTESİ

# KISALTMA LİSTESİ

**G.T.Ü.** **:** Gebze Teknik Üniversitesi
**DIGITS** **:** Derin Öğrenme GPU Eğitim Sistemi (Deep Learning GPU Training System)
**REST** **:** Temsili Durum Transferi (Representational State Transfer)
**API** **:** Uygulama Programlama Arayüzü (Application Programming Interface)
**JSON** **:** JavaScript Nesnesi Gösterimi (JavaScript Object Notation)

# ÖZET

Projenin amacı iki boyutlu görüntü üzerinde iskelet tahmini yapmaktır.

Veri artırma işlemi görüntüdeki kişinin her merkez eklemine en fazla 16 piksel uzaktaki komşu piksele uygulanmıştır. Bu komşu görüntüler, eklem merkezinden uzaklığı tanımlayan sınıflardır. Her komşu ve merkez piksel için, ilk 2B görüntüden 16x16, 32x32, 64x64 ve 128x128 çözünürlüklü resimler kırpılmıştır. Bu alt imgeler, sinir ağı modelini eğitmek için birbirine eklenmiştir. Her eklem için on beş derin öğrenme modeli elde edilmiştir.

Test aşamasında test görüntüsünün her pikseli için 15 ağdan tahmin alınarak 6 piksel ve altı uzaklık tahminleri doğru kabul edilmiş ve doğru olarak kabul edilen piksellerin lokasyon ortalamaları merkez olarak kabul edilmiştir. Ayaklarda ve görünmeyen noktalarda sapmalar gözlense de iskelet tahmininde başarılı bir sonuç elde edilmiştir.

## SUMMARY

This project aims to estimate human pose in a single 2D image.

Data augmentation operation is applied to at most 16 pixel far away neighbour pixel to every center joint of the person in the image. These neighbour images are the classes that define distance from the center of the joint. For every neighbour and center pixel, 16x16, 32x32, 64x64 and 128x128 resolution images are cropped from the first 2D image. These subimages are concatenated to each other to train the neural network model. Fifteen deep learning models are obtained for every joint.

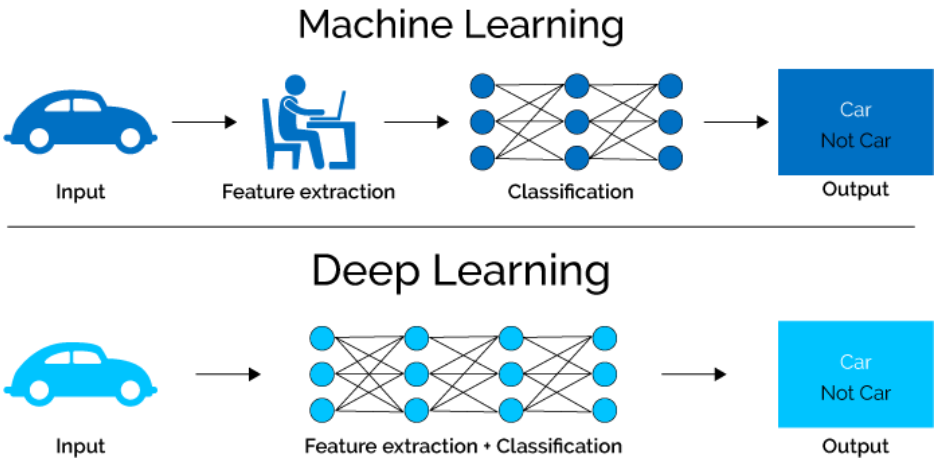
For each pixel of the test image during the test phase, estimates of distance from 6 pixels and below were taken correctly and the center of the locations of the correctly accepted pixels were considered as the center of the joint. Despite the deviation of the legs and invisible joints, the pose estimation can be considered successful.

# 1. ARKAPLAN

İki boyutlu bir görüntüden bir insanın iskelet pozisyonunu tahmin etmek bilgisayarla görü dünyasındaki zorlu görevlerden biridir. Sağlık alanında cerrahi operasyonlardan fizik terapiye, askeri alanda aksiyon tanımadan simülasyonlara, eğlence alanında hareket tanımlama ile oyunlardan insan-bilgisayar etkileşimine kadar çoğu alanda vazgeçilmez bir gereksinim olduğu kadar alanların çeşitliliğine göre de çeşitli zorluklara sahiptir.

Bilgisayarla görü ile anlamlandırma ve sınıflama, görüntü ve videolardan(görüntü akışları) özellik vektörleri oluşturarak( feature extraction ) bu vektörlerin makine öğrenmesi yöntemleriyle eğitilmesi ve bir model oluşturmasına dayalı idi[7].

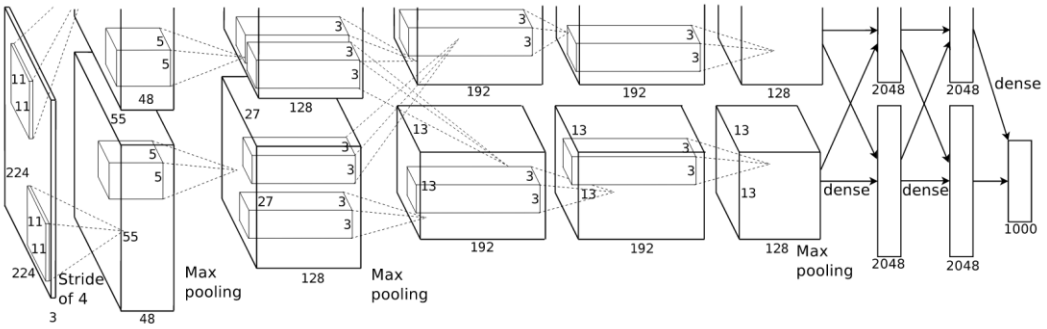
Makine öğrenimi tabanlı bilgisayarla görü yöntemleri başarı örnekleri gösterse de çevre ve kıyafet farklılığı, ışık, görüntünün kompleks olup olmaması, insanın sahip olabileceği pozisyonlar ve perspektif gibi özellikler başarı seviyesini düşürmektedir[1].




ŞEKİL 1.1 Makine öğrenmesi ve Derin Öğrenme arasındaki fark[8]

Derin öğrenme veya diğer adıyla sinir ağları, özellik çıkarılmadan, etiketlenmiş ham veriyi girdi olarak kabul ederek öğrenen yapay zeka alt dalıdır. Modele verilen veriyi iterasyonlar uygulayarak ve içindeki parametreleri düzenleyerek öğrenim sağlayan bir konsepttir. Derin öğrenme ile görüntü sınıflandırma ele alındığında, görüntüler convolution, pool, lineerizasyon ve aktivasyon katmanlarından oluşan bir ağda işlenmektedir. Bu katmanların parametrelerinin ayarlanması ise verilen eğitim görüntülerinin etiketlerine bağlıdır ve eğitim sırasında etiketlere göre geriye yayılım(back propagation) işlemi uygulanır. Geriye yayılım dışında çeşitli ağlar da mevcuttur.

Teknolojinin gelişmesine bağlı olarak sahip olduğumuz verilerin artışı ve donanımsal olarak güçlenen bilgisayarların kompleks matematiksel işlemleri daha hızlı yapabiliyor olması, paralel işlemci yapılarını baz alan ekran kartlarının (GPU) ortaya çıkması ve gelişmesi ile gerçeklenebilen sinir ağlarını üne kavuşturan en önemli çalışmalardan biri Krizhevsky ve diğerlerinin çalışması olan Alexnet'tir[2]. Alexnet modelinin 2012 yılında ImageNet[9] yarışmasında kayda değer bir başarı göstermesi bilgisayarla görü dünyasında büyük yankı uyandırmıştır.

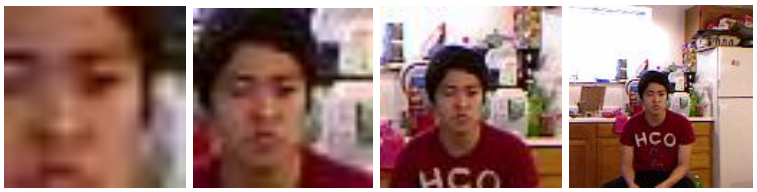



ŞEKİL 1.2 Alexnet Ağ Mimarisi[10]

İlerleyen yıllarda sınıflandırma yöntemlerinde başarıların artırılması ile[11,12,13,14] bilgisayarla görü problemlerinin çözüm başarıları artırılmıştır. Derin öğrenme ile iskelet pozisyonu tahmini üzerine yapılan çalışmalarda da başarılı sonuçlar elde edilmiştir[3,4].

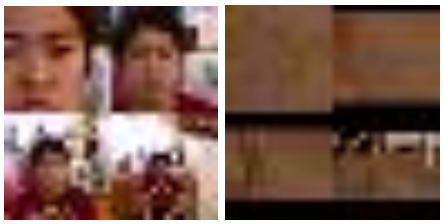
## 2. METOT

Önerilen metot daha önceki çalışmalarda[3,4] bulunan eklem öğrenim yönteminin yalın halde kullanılmasını önerir. İki boyutlu görüntüden ve eklem noktalarından, orjin başta olmak üzere 16 piksel uzaklığa kadar komşuluklardan yamalar(patch) alınır.



ŞEKİL 2.1 Belirlenen x ve y noktaları için alınan 4 farklı büyüklükte yamalar ( soldan sağa doğru 16x16, 32x32, 64x64, 128x128 )

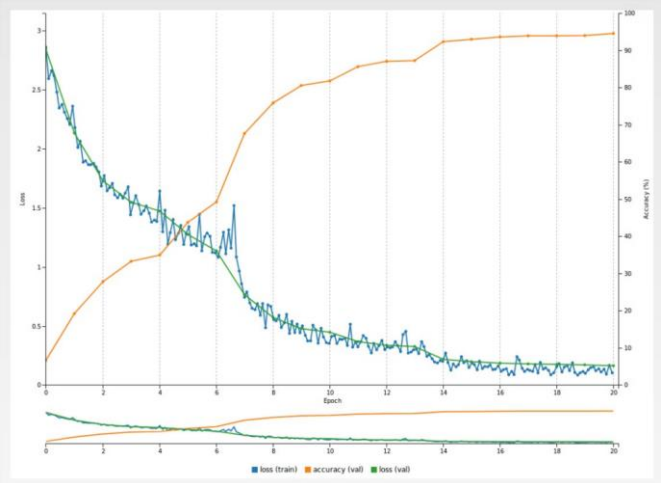
On beş eklemin XYZ eksenleri verisetinde önerilen dönüşüm fonksiyonu ile görüntü üstündeki XY piksel değerlerine dönüştürülürür. Dönüştürülen XY görüntü lokasyonu orjin kabul edilerek Öklid(L2) uzaklıklıklarına[6] göre çevresindeki pikseller sınıflandırılır ve her bir piksel merkezde kalacak şekilde yamalar alınır. Yamalar alınırken daha önceki çalışmalarda kullanılmış bir yöntem olarak[5], ŞEKİL 2.1 ve ŞEKİL 2.2'de görüldüğü gibi 16x16, 32x32, 64x64 ve 128x128 piksel çerçeveler 16x16 boyutuna indirgenir ve birleştirilir. Yamalar görüntünün köşe noktalarından alındığında, ŞEKİL 2.2'de görüleceği üzere alınan yama siyah pikseller ile doldurulur. Örnek vermek gerekirse 320x240 piksellik bir görüntüde sağ ayak x=290 y=200 noktasında ise alınacak 128x128 yama siyah noktalar ile doldurulmazsa görüntü elde edilemeyecektir. Bu sebeple siyah piksellerler doldurma işlemi gerçekleştirilir. Oluşan bu veri seti ile sinir ağı eğitilir. Sinir ağı olarak Alexnet[2], veriseti olarak Cornell CAD-60[5] veriseti kullanılmıştır. Veriseti, 4 kişinin 5 farklı mekanda gerçekleştirdiği 12 farklı eylemi, iskelet noktaları ile beraber içeren bir video verisetidir. Veriset içindeki 60 farklı videodan birer çerçeve(frame) elde edilip çerçevedeki kişinin 15 farklı eklemi için bahsedilen parçalanma işlemi tekrarlanmaktadır. Sonucunda 15 farklı eklem için ayrı veriseti oluşturulmuştur. Her veriseti seksen bine yakın görüntüyü içermektedir. Her eklem için ayrı bir ağ eğitilerek eklemlerin öğrenimi amaçlanmaktadır.



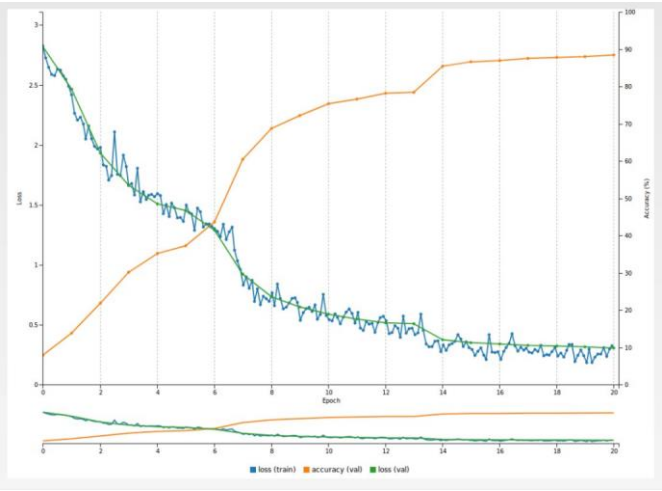
ŞEKİL 2.2 Farklı büyüklükte alınan görüntülerin eğitime verilmek üzere birleştirilmesi

Nvidia Digits ile arayüz yardımıyla Alexnet modelinde eğitilen ağlar, Digits REST API yardımıyla erişilerek test edilecek ve ilk 5 tahminin en iyisi JSON formatında alınacaktır. JSON formatında alınan veri anlamlı şekilde parçalanacak ve en iyi tahminler içinde 6 piksel ve daha yakın olarak tahmin edilen piksellerin lokasyon ortalaması tahmin edilen noktayı oluşturmaktadır.

# 3. DENEYLER VE BULGULAR




ŞEKİL 3.1 Baş için eğitilen ağın başarım oranı.




ŞEKİL 3.2 Sağ ayak için eğitilen ağın başarım oranı.

Eğitim Nvidia Digits[15] ile, caffe[16] dilinde %15 test, %20 geçerlilik datası ayırılarak gerçeklendi. GTX970 ile her eklemin eğitimi 2 saat sürdü. Dört farklı perspektifi birleştirmeden yapılan 16x16 ve 32x32'lik görüntü girdisine sahip olan ağlar, %99.5 başarım oranlarına çıktı. Bu ağlar test edildiğinde ağın ezberlediği anlaşıldı. Ağ sonuçlarında 600 test verisi ile test yapılıp analiz edildiğinde 200 civarında çok uzak gelen değerler(2 piksel uzaklıkta demesi gerekirken 14 piksel uzakta demesi gibi ) ve geri kalan datada da %100 oranlı başarım(2 piksel olması gereken test görüntüsünü %100 ihtimalle 2 piksel olarak bilmesi ve en iyi beş tahminden geri kalanların %0 ihtimale sahip olması) elde edildi. Test aşamasında başarısız olan bu ağlar yerine önerilen metoda geçildi. Başarım oranları her eklem için %90 dolaylarında olmakla beraber, maksimum başarım oranı ŞEKİL 3.1'de görülebileceği gibi baş ağında %94 ile, en az ise ŞEKİL 3.2'de görülebileceği gibi %88 ile sağ ayak ağında elde edildi. Bahsi geçen değerler validasyon verisi üzerinden elde edilen sonuçlardır, turuncu çizgi ile belirtilmişlerdir.

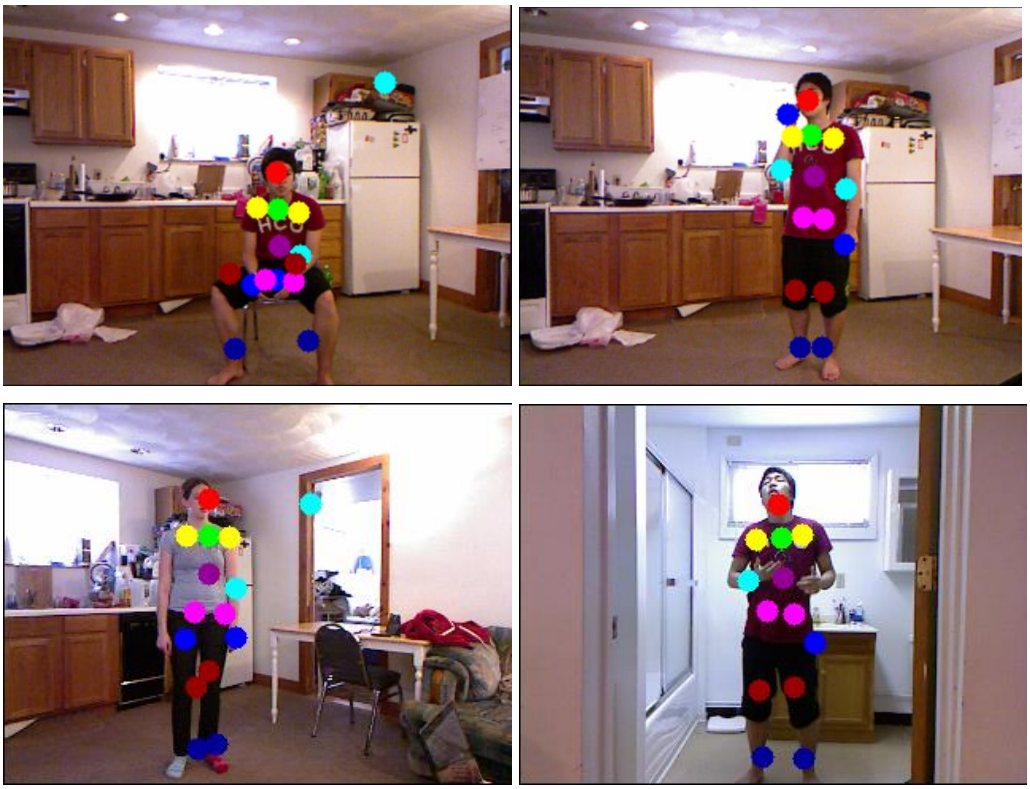
| Eklem | Test verisinde başarım oranı |
|---|---|
| **Baş** | % 0.778 |
| **Boyun** | % 0.688 |
| **Sağ El** | % 0.478 |
| **Sol El** | % 0.467 |
| **Sağ Omuz** | % 0.663 |
| **Sol Omuz** | % 0.678 |
| **Sağ Dirsek** | % 0.479 |
| **Sol Dirsek** | % 0.496 |
| **Sağ Kalça** | % 0.643 |
| **Sol Kalça** | % 0.687 |
| **Sağ Diz** | % 0.881 |
| **Sol Diz** | % 0.875 |
| **Sağ Ayak** | % 0.898 |
| **Sol Ayak** | % 0.8 |
| **Gövde** | % 0.701 |

TABLO 3.1 Eğitilen ağların test verisine göre başarım oranları

Test kısmında, CAD-60 verisetinin içerisinde bulunan görüntüler kullanıldı. Görüntülerde kişinin vücut iskeletinin aynı durmaması için aktivitenin eğitime verilen görüntüsünden 100 çerçeve sonraki görüntüler kullanıldı. Verisetinin elde edildiği Microsoft Kinect [17] cihazının 30 fps(saniyedeki çerçeve sayısı) olduğu düşünülürse 3.3 saniye sonrasına denk geldiği söylenebilir.

Test aşaması için seçilen görüntüler, eğitime verilir gibi eklem noktalarından kırpılarak 5400 dolaylarında görüntü elde edilmiştir. Test görüntüleri, piksel uzaklıkları bilindiğinden dolayı tahmin sonuçlarında karşılaştırılabilmektedir.

Olması gerekenden 1 piksel uzak veya yakın olarak veya tam olarak tahmin edilen sonuçlar doğru kabul edilmiştir. Test sonucunda, TABLO 3.3'de görülebileceği üzere el ve dirseklerde %50 gibi düşük bir başarım oranı, genel olarak eklemlerde %80 dolaylarında bir başarım oranı gözlenmektedir. El ve dirsek ağlarındaki başarım oranının düşük olma sebeplerinin el ve dirseklerin diğer eklemlere göre daha çeşitli yerlerde bulunabilmesi, bu eklemlerin diğer eklemlerin önüne geçebiliyor olması ve başarım oranı hesaplanırken 1 piksel gibi küçük bir uzaklık değerinin alınmasıdır. Uzaklık değeri 5 piksel değerine kadar genişletilebilir. Gözleri sağlıklı gören bir insanın şaşırma payının 320x240 görüntülerde maksimum 5 piksel olabileceği tahmin edilirse, başarım oranının artırılabileceği fakat tahmin aralığı dar tutulduğundan dolayı başarım oranının düşük bir yüzdede kaldığı düşünülebilir.



ŞEKİL 3.3 Test görüntülerinde elde edilen tahminler

Test verileri kırpma işlemi olmadan, her pikseli için bir görüntü oluşturularak modele verildiğinde ve 6 piksel ve daha yakın olarak tahmin edilen piksellerin lokasyonlarının ortalaması alınarak görselleştirildiğinde ŞEKİL 3.3'deki görüntüler elde edilir. ŞEKİL 3.3'de gösterildiği üzere bazı durumlarda eklemlerde hafif kayma olmuş, sağ dirsek ise iki görüntüde büyük bir hata ile anlamsız bir yerde tahmin edilmiştir. Sapmaların genel olmadığı düşünülürse iskelet uygun kabul edilecek düzeyde tahmin edilmiştir denilebilir. Sapmaların olmasında, 6 piksel ve daha yakında tahmin edilen piksellerin lokasyon ortalaması alınması, sınırlara yakın görüntülerde siyah pikseller ile görüntüyü çerçeveleme ve 64x64, 128x128 gibi çerçevelere tamamlama işlemlerinin sebep olduğu düşünülmektedir. Ayrıca 320x240 veriseti görüntülerinde 16x16 olarak alınan görüntü genel görüntüye göre yeterli bir orana sahip sahip olsa da eklemlerin analizi bazında düşünüldüğünde yeterli ayrıntıyı kapsamadığı ortaya çıkmaktadır.

# 4. SONUÇLAR

Gerçekleştirilen projede, Karakoç ve diğerlerinin[7] çalışmasından ilham alınarak eklemlerin x ve y noktalarının çeşitli görüntüleri birleştirilmiş ve eğitim gerçeklenmiştir. CAD-60 veriseti kullanılan çalışmada, aktivite videolarındaki düşük çözünürlük, sınır noktalarına uygulanan siyah pixel ile görüntü doldurma işlemi, tahmin gerçekleştirilen piksellerden 6 piksel ve aşağısı yakınlıkta olan piksellerin lokasyon ortalaması alınması, sonucu etkileyen en önemli noktalardır. Ayrıca, sağ veya sol elin başın arkasında kaldığı veya bazı eklemlerin görüntüde bulunmadığı durumlarda(kişiye yandan bakılan bir görüntü gibi) iskeletin eksik olarak tahmin edilmesi söz konusudur. Bahsi edilen başarısız konular dışında derin öğrenme tabanlı iskelet tahmin sistemi, test verilerinde başarılı sayılabilecek bir sonuca ulaşmıştır.

Çalışmanın test verisi üzerinde ve çeşitli görüntülerde daha iyi çalışması amacıyla çeşitli iyileştirmeler gerçekleştirilebilir. Veriseti daha çok çeşitte insanların bulunduğu, yüksek çözünürlüklü bir veriseti ile değiştirilebilir. Bu şekilde az kişi ile

öğrenme ve az ayrıntıyı ele alma gibi başarımı aşağı çeken parametreler ortadan kaldırılabilir. Bir buçuk saat süren test aşamasını minimize etmek amacıyla Nvidia Digits'in sağladığı REST API arayüzü yerine modelin kendisine sorgu yapılarak Digits ortamına harcanan görüntülerin yüklenme süresi ortadan kaldırılabilir. Görüntünün her pikseli üzerine gerçekleştirilen kırpma işlemi yerine, çeşitli piksel değerlerinde atlamalar yapılarak görüntünün yüzde biri test edilir ve ardından başarılı olarak kabul edilen pikseller üzerinden devam edilebilir. Başarı oranınn artması amacıyla, başarılı bulunduğu kabul edilen bir eklemin diğer eklemler aranırken merkez baz alınması sağlanabilir. Bu şekilde gövde ve dizler gibi başın altında kalan eklemler aranırken, ayakta duran bir kişi için görüntünün baş ekleminden aşağısı aranması sağlanır. Tahminlerin iterasyonu artırılarak başarı oranının artırılması da mümkündür. Çeşitli ağlarda eğitim yapılarak daha uygun bir ağ bulunabilir veya girdiye çeşitli işlemler uygulanabilir. Sonraki çalışmalar için daha iyi sonuçların elde edilmesi mümkündür.

# KAYNAKLAR


[1] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *European Conference on Computer Vision*. Springer International Publishing, 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[3] Xiaochuan Fan, Kang Zheng, Yuewei Lin, Song Wang. "Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation." *The IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[4] Tompson, Jonathan et al. "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation." *NIPS* , 2014.

[5] Computer Science Department, Cornell University. Cornell Activity Datasets: CAD-60 , http://pr.cs.cornell.edu/humanactivities/data.php, [Ziyaret Tarihi: 17 Aralık 2017].

[6] Öklid uzaklığı, Wikipeda page, https://en.wikipedia.org/wiki/Euclidean_distance, [Ziyaret Tarihi: 17 Aralık 2017].
[7] Necmeddin Said Karakoç, Şamil Karahan, Yusuf Sinan Akgül. "Deep learning based estimation of the eye pupil center by using image patch classification". *Signal Processing and Communications Applications Conference*, 2017.
[8] https://www.slideshare.net/datascienceth/machine-learning-in-image-processing , [Ziyaret Tarihi: 17 Aralık 2017].
[9] ImageNet veriseti ve yarışması, http://www.image-net.org/ [Ziyaret Tarihi: 17 Aralık 2017].
[10] Makine öğrenmesi ve derin öğrenme arasındaki fark, https://cdn-images-1.medium.com/max/1600/1*SVMH0_mLM9gH3miMPSsAgw.png , [Ziyaret Tarihi: 17 Aralık 2017].
[11] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.
[12] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint *arXiv*:1409.1556 (2014).
[13] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
[14] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[15] Nvidia Digits Framework, https://developer.nvidia.com/digits , [Ziyaret Tarihi: 17 Aralık 2017].
[16] Caffe Deep Learning Framework http://caffe.berkeleyvision.org/ , [Ziyaret Tarihi: 17 Aralık 2017].
[17] Microsoft Kinect, https://msdn.microsoft.com/en-us/library/jj131033.aspx , [Ziyaret Tarihi: 17 Aralık 2017].