**Derin Sinir Ağlarıyla Osmanlıca Optik Karakter Tanıma (OCR)**

**Çalışmanın Amacı**

Osmanlıca, 13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğu'nda kullanılan ve Arap alfabesi tabanlı bir yazı dilidir. Osmanlıca ile yazılmış milyonlarca belge, kitap, dergi ve resmi kayıt günümüzde arşivlerde bulunmaktadır. Ancak bu belgelerin çok küçük bir kısmı dijital ortama aktarılmıştır. Geleneksel Osmanlıca OCR çalışmaları sınırlıdır ve mevcut sistemler tatmin edici sonuçlar verememektedir.

Bu çalışmanın temel amacı, matbu nesih hattı ile basılmış Osmanlıca belgelerde karakter tanımada yüksek başarı sağlayan derin sinir ağları tabanlı bir OCR modeli geliştirmektir. Çalışma kapsamında, Convolutional Neural Networks (CNN) ve Long Short-Term Memory (LSTM) birleşimi bir model kullanılarak, Osmanlıca optik karakter tanıma süreci geliştirilmiş ve mevcut OCR sistemleri ile karşılaştırılmıştır.

**Yöntem ve Kullanılan Veri Setleri**

Osmanlıca karakter tanıma için derin öğrenme tabanlı bir OCR modeli geliştirilmiştir. Modelin eğitimi için üç farklı veri kümesi oluşturulmuştur:

1. **Orijinal Veri Seti:** 1.000 sayfa Osmanlıca belge içermektedir.
2. **Sentetik Veri Seti:** 23.000 sayfa, dijital ortama aktarılan ve yapay olarak üretilen Osmanlıca metinlerden oluşmaktadır.
3. **Hibrit Veri Seti:** Orijinal ve sentetik verilerin birleştirilmesiyle oluşturulmuştur.

Test aşamasında 21 farklı sayfadan oluşan bir test veri kümesi kullanılmıştır. Modelin performansı; karakter tanıma, bağlı karakter tanıma ve kelime tanıma doğruluğu açısından değerlendirilmiş ve Google Docs, Abby FineReader, Tesseract’ın Arapça ve Farsça modelleri ile Miletos OCR aracıyla karşılaştırılmıştır.

OCR sürecinde metinlerin doğruluk analizini yapmak için üç farklı veri türü kullanılmıştır:

* **Ham metin:** OCR çıktısının doğrudan değerlendirilmesi.
* **Normalize metin:** Metinlerdeki yazım hatalarının düzeltilmesi.
* **Bitişik metin:** Boşlukların kaldırılarak kelimelerin bitiştirilmesiyle elde edilen metin.

OCR doğruluk oranlarını belirlemek için **karakter, bağlı karakter katarı ve kelime bazlı doğruluk oranları** hesaplanmıştır. Model çıktıları normalizasyon algoritmalarıyla düzeltilmiş ve hatalar silme, ekleme ve değiştirme türlerine ayrılarak analiz edilmiştir.

**Osmanlıca OCR’da Kullanılan Model**

Bu çalışmada **Convolutional Recurrent Neural Network (CRNN)** mimarisi kullanılmıştır. Modelin bileşenleri şu şekildedir:

1. **CNN (Convolutional Neural Network):** Görüntü verisinden öznitelik çıkarma işlemi için kullanılır.
2. **RNN (Recurrent Neural Network) ve LSTM:** OCR sürecinde karakterlerin sıralı verilerle ilişkilendirilmesini sağlar.
3. **CTC (Connectionist Temporal Classification):** Modelin çıktılarındaki olasılık dağılımını düzenleyerek harfleri doğru sıraya koymaya yardımcı olur.

Modelin eğitiminde:

* **Öğrenme oranı (learning rate):** 0.002
* **Momentum:** 0.5
* **Epoch sayısı:** 3.000.000 olarak belirlenmiştir.

**Sonuçlar ve Karşılaştırmalar**

Çalışma kapsamında geliştirilen Osmanlıca OCR modeli, mevcut OCR sistemleriyle karşılaştırıldığında belirgin bir üstünlük göstermiştir. **Karakter, bağlı karakter ve kelime tanıma doğruluk oranları** aşağıdaki gibi hesaplanmıştır:

| **Model** | **Ham Metin (%)** | **Normalize Metin (%)** | **Bitişik Metin (%)** |
| --- | --- | --- | --- |
| Osmanlıca Hibrit | **88.86** | **96.12** | **97.37** |
| Osmanlıca Orijinal | 87.73 | 94.87 | 96.16 |
| Osmanlıca Sentetik | 73.16 | 77.64 | 78.10 |
| Google Docs | 83.86 | 92.02 | 91.43 |
| Abby FineReader | 71.98 | 80.19 | 81.05 |
| Tesseract (Arapça) | 76.92 | 82.37 | 81.27 |
| Tesseract (Farsça) | 75.30 | 83.85 | 83.48 |
| Miletos | 75.76 | 86.46 | 86.88 |

**Öne Çıkan Sonuçlar**

* **Osmanlıca Hibrit modeli, tüm testlerde en yüksek doğruluk oranını yakalamıştır.** Karakter tanımada %97.37, kelime tanımada %66.45 doğruluk oranına ulaşmıştır.
* **Google Docs ve Tesseract gibi popüler OCR araçlarına kıyasla %4-5 daha iyi karakter tanıma doğruluk oranları elde edilmiştir.**
* **Sentetik veri setiyle eğitilen model, orijinal veriye göre düşük performans göstermiştir.** Ancak hibrit veri kullanıldığında sentetik verinin katkı sağladığı görülmüştür.
* **Osmanlıca’da kelime tanıma oranları düşüktür, ancak bağlı karakter tanıma başarısı oldukça yüksektir.** Modelin en büyük zorluklarından biri kelimeler arası boşlukların doğru tespit edilmesi olmuştur.

**Sonuç ve Çalışmanın Katkısı**

Bu çalışma, Osmanlıca optik karakter tanıma alanında yapılan önceki çalışmalardan daha yüksek başarı oranına sahip bir model sunmaktadır. **CNN ve LSTM birleşimi bir derin sinir ağı mimarisi kullanılarak Osmanlıca matbu nesih fontunda %97'ye varan karakter tanıma doğruluk oranına ulaşılmıştır.**

Çalışmanın getirdiği yenilikler:

* **Geliştirilen model, Google Docs, Abby FineReader, Tesseract gibi popüler OCR araçlarına kıyasla daha yüksek doğruluk sağlamaktadır.**
* **Osmanlıca için derin öğrenme tabanlı bir OCR modeli önerilmiştir.** Bu model, Osmanlıca belgelerin dijital ortama aktarılmasını hızlandırarak arşivlerde bulunan milyonlarca belgenin erişilebilir hale getirilmesini sağlayabilir.
* **Osmanlıca karakterleri, bağlı karakter katarları ve kelime frekanslarının ayrıntılı analizleri sunulmuştur.** Bu analizler, OCR ve Osmanlıca Doğal Dil İşleme (NLP) alanındaki diğer çalışmalara katkı sağlayabilir.
* **Geliştirilen model Osmanlica.com adresinde kullanıma sunulmuştur.** Böylece kullanıcılar Osmanlıca belgeleri tarayarak metne dönüştürebilirler.

Bu araştırma, Osmanlıca OCR'nin gelişmesine büyük katkı sağlayarak tarih, dilbilim ve arşivcilik alanlarında önemli bir yenilik sunmaktadır. Gelecekteki çalışmalar, Osmanlıca el yazısı OCR'sini geliştirmeye ve modelin el yazmalarında da yüksek doğruluk sağlamasına odaklanabilir.