SISTEMA MÓVEL PARA ÁUDIO-TRANSCRIÇÃO DE ESCRITA CURSIVA

Trabalho de Conclusão de Curso
Engenharia da Computação

Felipe Mendonça Gouveia
Orientador: Prof. Dr. Byron Leite Dantas Bezerra

UNIVERSIDADE
DE PERNAMBUCO

Felipe Mendonça Gouveia

SISTEMA MÓVEL PARA ÁUDIO-TRANSCRIÇÃO DE ESCRITA CURSIVA

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco - Universidade de Pernambuco

Universidade de Pernambuco

Escola Politécnica de Pernambuco

Graduação em Engenharia de Computação

Orientador: Prof. Dr. Byron Leite Dantas Bezerra

Recife - PE, Brasil

12 de abril de 2015

Declaro que revisei o Trabalho de Conclusão de Curso sob o título *"SISTEMA MÓVEL PARA ÁUDIO-TRANSCRIÇÃO DE ESCRITA CURSIVA"*, de autoria de *Felipe Mendonça Gouveia*, e que estou de acordo com a entrega do mesmo.

Recife, _____ / _____ / _____

_____

Prof. Dr. Byron Leite Dantas Bezerra
Orientador

The right man in the wrong place can make all the difference in the world.
(GMan, Half-Life)

Agradecimentos

De todas as parte difíceis deste trabalho, talvez essa tenha sido a que me deu mais trabalho, não por não saber a quem agradecer, mas por não saber como expressar em palavras tudo aquilo que tenho para dizer para todas as pessoas que me ajudaram ao longo desta jornada. Gostaria então de agradecer a minha mãe e irmão, por sempre estarem ao meu lado me apoiando, minhas avós, meu avô, tias, tios e primas, e a toda a minha família. Todos os meu colegas que fiz ao longo da faculdade e durante meu estágio, meu orientador e meus professores. E a Bené, meu cachorro, por não ter comido este nem nenhum outro dos meus trabalhos.

Resumo

A escrita é uma das principais formas de comunicação e de troca de informações da humanidade. Possibilitar que as pessoas que tenham algum tipo de deficiência visual possam ter acesso a esse tipo de informações é o objetivo deste trabalho. Nele apresentamos um sistema ponto a ponto, que a partir da captura de uma imagem utilizando-se a câmera de um smartphone. O sistema processa a imagem capturada, reconhece o texto presente nesta utilizando técnicas do estado da arte, e sintetiza o resultado em formato de áudio para o usuário.

Palavras-chave: Escrita cursiva, síntese de áudio, processamento de imagem, redes neurais, LSTM.

Abstract

Handwriting is one of the oldest forms of communication and information exchange of mankind. Allow people to have access to that information is the objective of this work. In this work we present an end to end system, capable of capture an image from a smartphone camera. With the captured image we process it, use state of art recognition algorithms and synthesize the result to audio, giving it to the final user.
Keywords: Handwritten recognition, audio synthesizes, image processing, neural network, LSTM.

Lista de ilustrações

Lista de abreviaturas e siglas

LSTM	*Long Short-Term Memory*
RNA	*Redes Neurais Artificiais*
ITU	*International Telecomunication Union*
RGB	*Red, Green, Blue*
HSV	*Hue, Saturation, Value*
MDRNN	*Multidimensional recurrent neural network*
CTC	*Connectionist Temporal Classification*

Sumário

1 Introdução

A escrita é uma das formas de comunicação mais antigas, seu princípio data de por volta do quarto milênio antes de Cristo, mas se considerarmos as pinturas e desenhos rupestres teremos uma data ainda mais antiga. Com o passar do tempo e crescimento da sociedade, servindo como ponto de manutenção de informações que precisavam ser mantidas, como registros comerciais, seja para comunicação a distância através de cartas e telegramas, a escrita passou a cada vez mais fazer parte da vida da sociedade como um todo.

A cada ano mais e mais informações são produzidas de forma escrita, principalmente de forma cursiva, como receitas médicas, anotações de aula, cheques bancários, cartas, entre muitas outras formas. Fazendo com que a escrita seja parte fundamental da sociedade moderna. No entanto essa informação não está presente e acessível para todos os indivíduos. Segundo o Relatório Mundial sobre deficiências (WHO, 2015) cerca de 45 milhões de pessoas no mundo são cegas, e outras 135 milhões sofrem de alguma dificuldade ou problema visual, que os impede ou dificulta bastante de ter acesso a um conjunto imenso de informação que nos é veiculado de forma visual. Uma dessas informações é a escrita.

Um outro elemento que vem cada vez mais se tornando parte da sociedade moderna são os smartphones. Como é apontado pelo *International Telecomunication Union* (ITU) (ITU, 2015) a quantidade de dispositivos móveis ativos no mundo está por volta de 6,8 bilhões, com o acesso a internet nesses dispositivos sendo registrado em cerca de 2,3 bilhões.

O crescimento na adoção de smartphones por parte da população, faz com que cada vez mais se tenha investimento nesta tecnologia, possibilitando o desenvolvimento de smartphones com um maior poder computacional. Com esses avanços e a preocupação de possibilitar uma qualidade de vida melhor para toda a população notou-se a possibilidade do desenvolvimento de um sistema móvel capaz de reconhecer textos manuscritos e transcrevê-los para áudio.

1.1 Trabalhos Relacionados

Processamento de documentos capturados através de foto não é uma área recente, e vem sendo discutido a um bom tempo na academia. Um survey sobre o assunto é descrito por Liang em (LIANG; DOERMANN; LI, 2005), demonstrando os principais desafios na área, como baixa resolução dos dispositivos, e problemas de *dewarp* das imagens. Ferreia (FERREIRA; GARIN; GOSSELIN, 2005) nos apresenta uma proposta de sistema para pré-processamento e detecção de texto em imagens obtidas a partir de câmera.

Sistemas para auxílio de pessoas com dificuldades visuais também vem sendo bastante discutido ultimamente, Ezaki (EZAKI et al., 2005) nos apresenta um sistema similar sendo que o usuário tem a necessidade de utilizar um computador, não sendo nada portátil. Plotz em (PLÖTZ; THURAU; FINK, 2008) nos apresenta um sistema capaz de realizar o processamento de imagens a partir de quadro branco, sendo que sua técnica se baseia no uso de reconhecimento de escrita online.

1.2 Objetivos

O objetivo principal deste trabalho é o de desenvolver um sistema computacional para auxiliar pessoas que sofrem de dificuldades visual a conseguirem ter acesso a textos escritos, em especial aqueles compostos por partes manuscritas.

1.2.1 Objetivos Específicos

- Criação de uma base de palavras em português para o treinamento de sistemas de reconhecimento de escrita.
- Desenvolvimento de um aplicativo para dispositivos móveis capaz de capturar uma imagem, processá-la, e gerar como saída a áudio transcrição do texto contido na imagem.

1.3 Estrutura da Monografia

Este trabalho está organizado em cinco capítulos. No capítulo 2 é apresentado a discussão teórica acerca de processamento digital de imagens, formas de reconhecimento de escrita e síntese de áudio, bem como é feito um levantamento de trabalhos relacionados. No capítulo 3, discutimos como foi realizado processo de desenvolvimento do sistema. Continuando, no capítulo 4, discutimos os resultados alcançados, o comportamento do sistema bem como a criação de uma base de palavras em português. Por fim no capítulo 5, são apresentadas nossas conclusões e indicações de trabalhos futuros.

2 Referencial Teórico

Este capítulo tem por objetivo apresentar o conteúdo necessário para melhor entendimento das partes seguintes deste trabalho. Aqui iremos explanar conceitos tais como processamento de imagens, redes neurais, síntese de voz e reconhecimento de escrita.

2.1 Processamento Digital de Imagens

O Conjunto de técnicas para descrição e manipulação digital de imagem é chamado de processamento digital de imagens. Uma imagem digital pode ser vista como uma matriz, onde cada posição desta é chamada de pixel e representa uma informação de cor. Este valor pode ser um valor booleano, indicando presença ou não de informação, utilizado para imagens preto e branco, um valor escalar simples entre 0 e 255 no caso de imagens em tons de cinza (algumas implementações utilizam um valor de ponto flutuante entre 0 e 1). Uma imagem pode ainda ser descrita como um vetor composto por 3 ou 4 componentes para imagens coloridas ou coloridas com transparência respectivamente, onde cada posição do vetor representa os canais verde, vermelho, azul e alfa, isto para quando se trata do espaço de cores RGB (*Red, Green and Blue*). Para outros espaços de cores, como o HSV (*Hue, Saturation and Value*) o vetor de cores descreve os valores de brilho, saturação e intensidade.

Um sistema de processamento digital de imagens possui normalmente as etapas mostradas na Figura 1 (GONZALEZ,), descritas nas próximas seções.

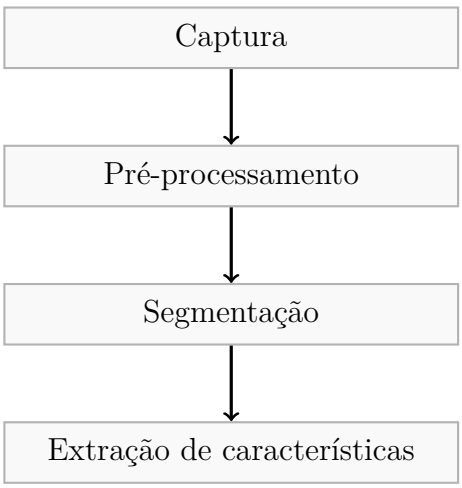
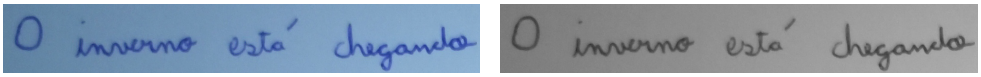


Figura 1 – Fluxo típico de um sistema de processamento digital de imagens.

2.1.1 Pré-processamento

A etapa de pré-processamento consiste em utilizar filtros para melhorar a qualidade da imagem para as etapas seguintes. Normalmente, é nesta etapa que se aplicam filtros de suavização e de eliminação de ruído, bem como é realizada a mudança no espaço de cores para aquele que melhor se adéqua ao problema em questão. Esses filtros em sua maioria são aplicados através do uso de uma janela deslizante sobre a imagem e a convolução desta com a região em que se encontra na imagem.

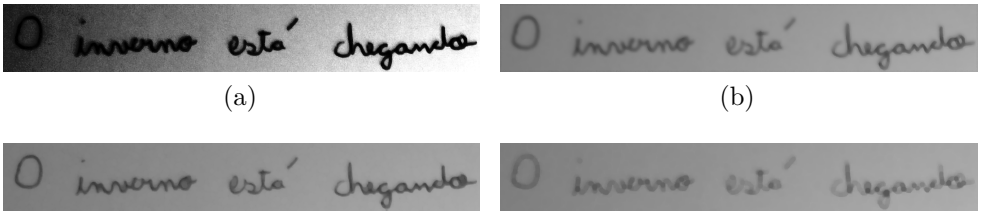
Uma etapa comum no processamento de documentos é a de conversão da imagem quando colorida para a escala de tons de cinza (STAMATOPOULOS et al., 2013) como mostrado na Figura 2.



(a) (b)

Figura 2 – Conversão imagem colorida 2a para tons de cinza 2b.

Com a imagem em tons de cinza a aplicação de várias operações se torna menos custosa. A aplicação destas operações tem por objetivo gerar uma versão da imagem que melhor se ajuste ao sistema desejado, como a equalização de histograma vide Figura 3a, filtro de remoção de ruído de Wiener vide Figura 3b, ou a aplicação de operadores morfológicos vide Figura 3c e Figura 3d (STAMATOPOULOS et al., 2013).



(a) (b)

(c) (d)

Figura 3 – Aplicação de vários filtros de processamento de imagens na Figura 2b. 3a equalização de histograma, 3b filtro de Wiener para remoção de ruído com janela tamanho 9×9, 3c operação de morfológica de abertura, 3d operação morfológica de fechamento.

2.1.2 Segmentação

O processo de extração de elementos da imagem com o objetivo de simplificá-la é chamado de segmentação.

Para o processo de extração de texto em uma imagem o que se objetiva é remover o plano de fundo e elementos não textuais para que fiquemos somente com o texto puro.

Uma técnicas bastante conhecida nesta área é a binarização, processo para transformar uma imagem cinza ou colorida em preto e branco, de Otsu (OTSU, 1975) que procura um valor que maximize a variância inter-classe de modo a se ter uma máxima separação entre fundo e objeto vide Figura 4. Embora muito utilizada a técnica de Otsu utiliza-se de um limiar calculado globalmente, ou seja, para toda a imagem, por consequência ele acaba não tendo resultados muito satisfatórios em imagens com uma grande variação de iluminação (NIBLACK, 1985). Neste cenário técnicas mais robustas e de aplicação local, ou seja, que se restringe a regiões específicas da imagem, tem alcançado resultados melhores do que técnicas globais vide (NIBLACK, 1985) e (SAUVOLA; PIETIKÄINEN, 2000).

Figura 4 – Binarização da imagem apresentada em 2b utilizando a técnica de Otsu.

Uma técnica bastante utilizada para a segmentação de documentos é a de Niblack (NIBLACK, 1985) e suas variações (SAUVOLA; PIETIKÄINEN, 2000). Essas técnicas buscam definir se um dado pixel ou região de pixels em uma imagem pertencem ao grupo de fundo ou objeto de acordo com regras pré estabelecidas. A técnica de Niblack segue a fórmula definida em 2.1, possuindo dois parâmetros básicos, a variável k que ajusta o filtro, e o tamanho da janela ao redor do pixel que será utilizada, uma janela do tamanho da imagem pode ser vista como um processo de segmentação global.

$$T = m + k * d \tag{2.1}$$

Onde, m é o valor médio dos pixeis que componham a janela ao redor do pixel que se deseja saber se é objeto ou fundo e d o desvio padrão das intensidades dos pixeis dentro da janela. T é o valor do limiar a ser calculado, sendo considerado o pixel pertencente ao objeto se tiver seu valor acima do limiar e não pertencente se tiver o seu valor abaixo do limiar. Uma aplicação da técnica de Niblack é apresentada na Figura 5.

Figura 5 – Binarização da imagem apresentada em 2b utilizando a técnica de Niblack.

Uma evolução da técnica de Niblack é a apresentada po Sauvola em (SAUVOLA; PIETIKÄINEN, 2000). Assim como na equação de Niblack o valor do limiar de cada pixel é computado segundo sua vizinhança, sendo que agora a equação utilizada é a presente na equação 2.2, onde, m é a média dos valores dos pixeis dentro da janela, k é um parâmetro,

s é o desvio padrão das intensidades dos pixeis na janela e R é um parâmetro para o desvio, normalmente utilizado como 128. O resultado da aplicação de Sauvola na Figura 2b é apresentado na Figura 6.

$$T_s = m + (1 - k \cdot (1 - s/R)) \tag{2.2}$$

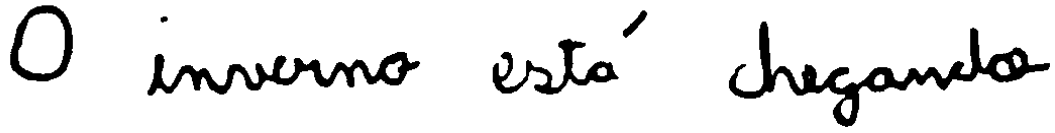


Figura 6 – Binarização da imagem apresentada em 2b utilizando a técnica de Sauvola.

2.2 Redes Neurais Artificiais

Em uma definição minimalista redes neurais artificiais são modelos computacionais que se baseiam no funcionamento do cérebro humano. De forma simples, uma rede neural é um conjunto de unidades computacionais simples interconectadas seguindo uma arquitetura definida, que após um processo de treinamento, processo este que visa otimizar as conexões entre os elementos da rede de forma a melhor resolver o problema proposto, apresenta uma solução com uma taxa de sucesso conhecida. Um dos modelos mais utilizados de redes neurais é o modelo conhecido como *multi layer perceptron*, ou perceptron de múltiplas camadas (MLP), esse modelo é uma evolução do modelo de Rosemblat (ROSENBLATT, 1958), e apresenta uma arquitetura semelhante a mostrada na Figura 7

2.2.1 LSTM

Embora tenham se mostrado bastante robustas para resolução de problemas de reconhecimento de padrões (ZHANG, 2000), redes neurais apresentam um problema conhecido como *vanishing gradient problem* (Figura 8) (HOCHREITER et al., 2001). Este problema se caracteriza pelo fato de a informação na rede se perder ao longo do tempo de treinamento, fazendo com que esta não consiga aprender certos tipos de problemas.

Para solucionar este problema, Hochreiter e Schmidhuber (HOCHREITER; SCHMIDHUBER, 1997) propuseram a técnica chamada de *Long Short-Term Memory* (LSTM), que foi estendia em 2001 por Gers (GERS, 2001) e depois melhorada por Gers em (GERS; SCHRAUDOLPH; SCHMIDHUBER, 2003). LSTM se baseia no conceito de células de memória, onde cada célula possui um conjunto de portas responsáveis pelo controle do fluxo da informação pelos neurônios da rede. Uma porta se comporta de forma similar a um neurônio, possuindo um conjunto de ligações de entrada e uma função de ativação, mas diferentemente de um neurônio as unidades de porta utilizam uma função de multiplicação

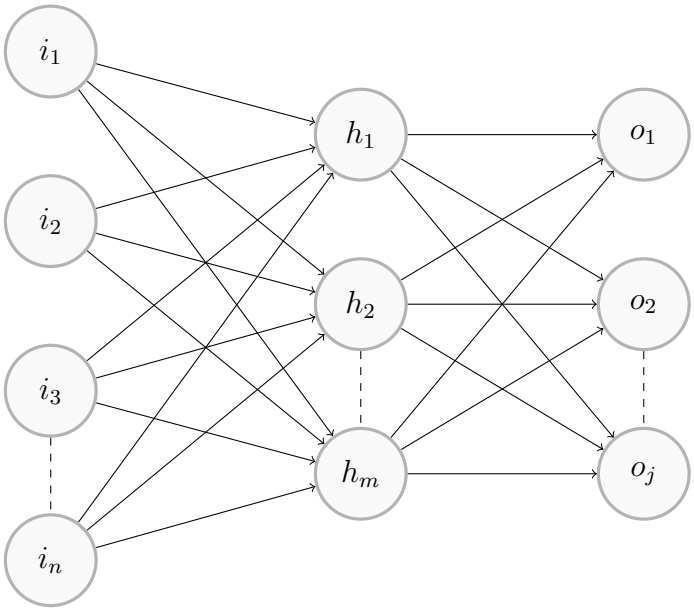


Figura 7 – MLP com n neurônios na camada de entrada, uma única camada de saída com m neurônios e j neurônios na camada de saída

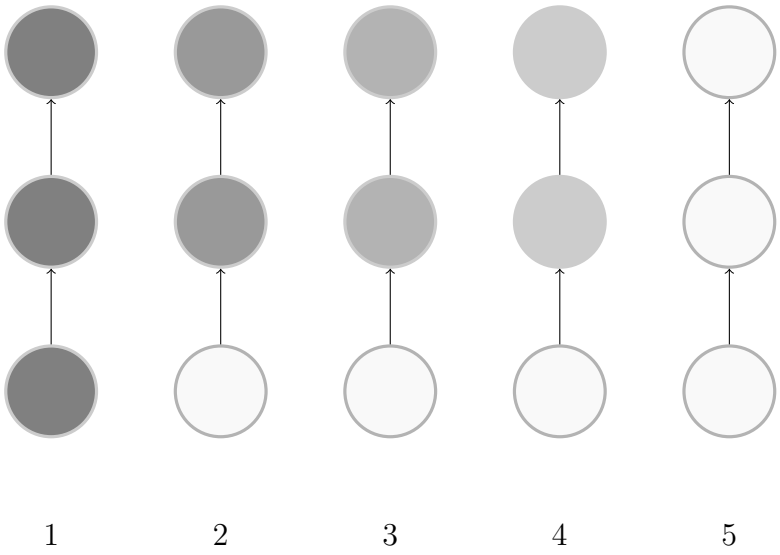


Figura 8 – Gradiente decaindo ao longo das épocas

para determinar o valor da informação. Uma célula LSTM possui 3 portas de controle, sendo uma responsável pela entrada de informação na célula, uma pela liberação da informação na célula, e uma terceira responsável por limpar a informação na célula como pode ser visto na Figura 9.

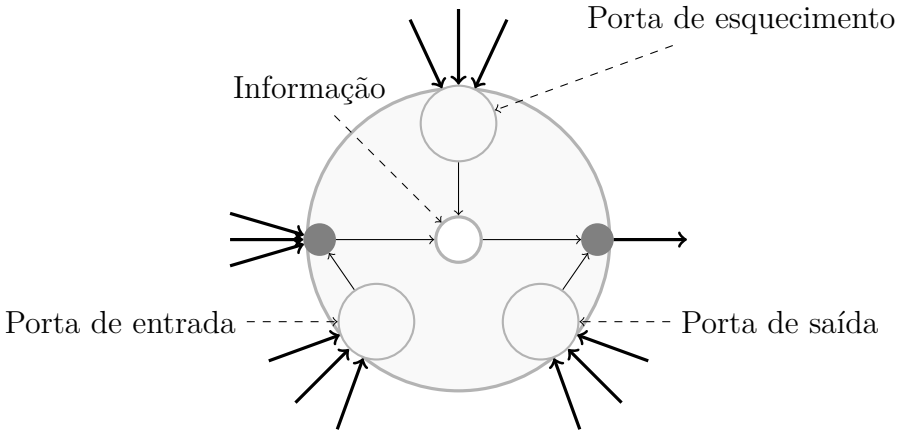


Figura 9 – Célula LSTM

Como pode-se notar ao se utilizar uma célula LSTM a quantidade de pesos e consequentemente de conexões aumenta em 3 vezes, aumentando também a complexidade computacional do modelo. Para resolver esse problema sem perda de desempenho Hochreiter (HOCHREITER; SCHMIDHUBER, 1997) propõe a ideia do uso de blocos, um bloco é um conjunto de células que compartilha as portas, diminuindo assim o custo computacional. Em comparação com uma rede neural MLP simples, a informação em uma rede LSTM tende a não se perder tão facilmente, podendo persistir na rede por várias épocas. A Figura 11 apresenta o fluxo ao longo do tempo de uma rede LSTM, a camada escondida desta usa a notação apresentada na Figura 10.

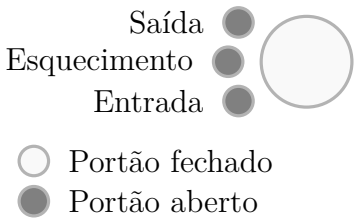


Figura 10 – Modelo LSTM.

Na Figura 11, ao ser apresentada uma informação para a rede esta só se propaga para a camada escondida caso a porta de entrada "permita", ou seja, seu estágio de ativação seja alcançado, estágio este alcançado na apresentação da segunda informação a rede. De modo análogo, uma informação só é propagada da camada LSTM atual para a próxima camada, caso a sua porta de saída seja acionada, situação que ocorre na apresentação

da quarta informação. Na apresentação da quinta informação para a rede a porta de esquecimento é ativada, fazendo com que a informação presente na célula seja esquecida, ou seja, seu valor seja zerado.

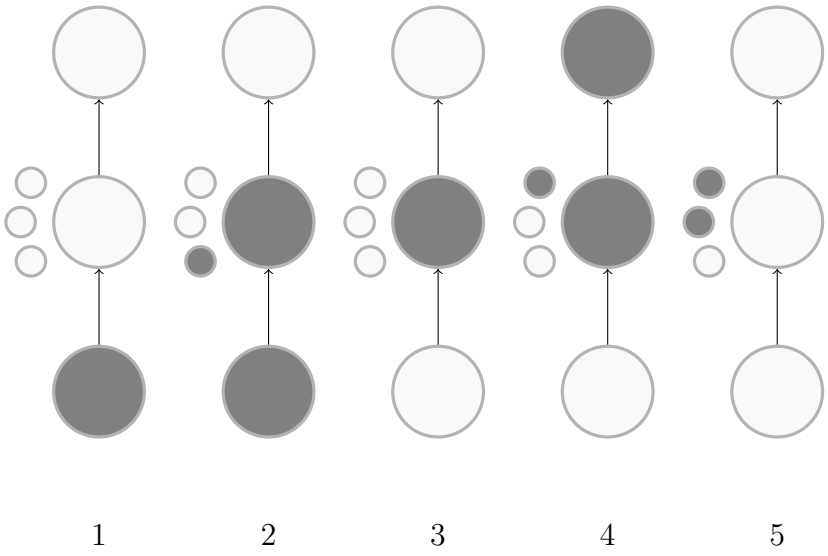


Figura 11 – Fluxo da informação ao longo das épocas em uma rede LSTM.

2.3 Reconhecimento de escrita

Reconhecimento de escrita é a área da computação que tem por objetivo dado uma entrada que pode ser uma imagem, ou uma sequência de movimentos de um marcador, por exemplo, reconhecer que naquela entrada existe informação textual, e dizer qual o texto que ali se encontra.

Existem dois tipos básicos de sistemas de reconhecimento de escrita (BUNKE, 2003). O sistema de reconhecimento online é aquele em que a se trabalha com a informação temporal de como a escrita foi desenvolvida, este tipo de sistema normalmente é realizado utilizando-se um quadro branco e um marcador com sensores para aquisiação dos dados (GRAVES et al., 2009), ou dispositivos compostos por tela de toque, como os antigos palmtops e os smartphones da linha Note da samsung (SAMSUNG, 2015).

O sistema offline é aquele em que a única fonte de informação para processamento e reconhecimento do texto é a imagem deste, tornando esta tarefa de caráter mais difícil (BUNKE; BENGIO; VINCIARELLI, 2004).

2.4 Síntese de Áudio

Síntese de áudio é o processo em que dado uma informação textual um sistema produz o áudio correspondente (DUTOIT, 1997). Uma das formas mais simples de se

produzir uma síntese de áudio é a de se pegar um conjunto de palavras, gravar cada uma separadamente e utilizar uma junção destes áudios individuais parar produzir uma saída desejada. No entanto esta técnica se mostra ineficiente, já que, o tamanho da base teria que ser compatível com a quantidade de palavras existentes na língua de destino, tornando o espaço necessário e o esforço para criação de tal modelo bastante desvantajoso. Como a forma de se representar textos escrito e por áudio varia, é necessário uma forma de representação para o áudio. Em O'Saughnessy (O'SHAUGHNESSY, 1987) mostra que cada linguagem possui um conjunto de fonemas, e que este varia entre 20 a 60 símbolos. Técnicas mais robustas de síntese de áudio se utilizam de fonemas para gerar seus resultados (LEMMETTY, 1999).

Os conceitos vistos neste capítulo embora distintos entre si, são necessários para o entendimento do sistema completo que será apresentado no próximo capítulo.

3 Sistema móvel para áudio-transcrição de escrita cursiva

O sistema desenvolvido neste trabalho é um sistema de visão computacional composto por cinco módulos principais a serem discutidos nas sub seções seguintes. Um esquema do projeto está apresentado na Figura 12. Optou-se por utilizar a plataforma Android para o desenvolvimento deste, pois pode ser considerada a plataforma mais popular entre a população (IDC, 2015), propiciando assim uma aplicação que atinge uma gama maior de indivíduos.

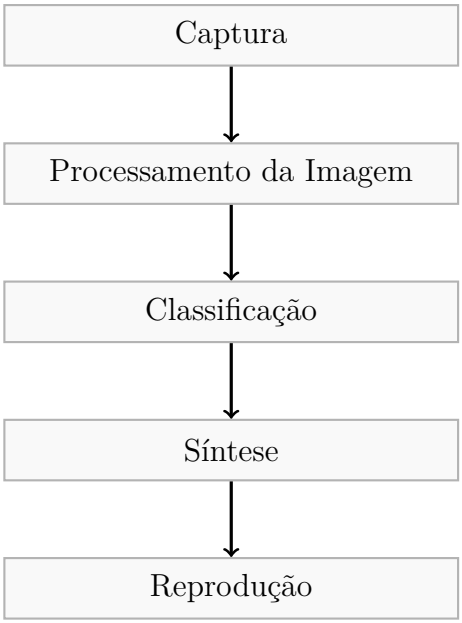


Figura 12 – Sistema móvel para áudio-transcrição de escrita cursiva.

3.1 Android

Android é o nome dado ao sistema móvel desenvolvido pelo grupo Android Inc em 2003 e posteriormente comprado pela Google em 2005. O sistema é baseado no kernel linux, e executa por cima deste uma máquina virtual java. Cada programa no android executa como sendo seu próprio usuário, com suas regiões de memória e dados separadas (LECHETA, 2013).

Um aplicação *android* consiste em uma ou mais *Activities*, cada *Activity* é uma tela da aplicação, serviços, provedores de conteúdo e receptores de *broadcast*. Um dos princípios do desenvolvimento para *android* é que cada aplicativo pode iniciar um determinado

componente de outro aplicação (ANDROID, 2015). O ciclo de uma aplicação funciona como uma pilha, ao ser instanciada uma *activity* assume o topo da pilha e inicia sua execução, ao ser requisitada uma nova *activity* o sistema coloca esta no topo da pilha e pausa a execução da anterior, quando a nova *activity* termina sua execução esta sai da pilha e a anterior volta a sua execução normalmente.

3.2 Captura

Ao iniciar a aplicação o usuário estará no módulo de captura, este é um dos módulos mais simples da aplicação e consiste apenas da tela apresentada na Figura 13, onde o usuário tem a opção de tirar uma nova foto, ou de escolher uma imagem presente no sistema do smartphone.

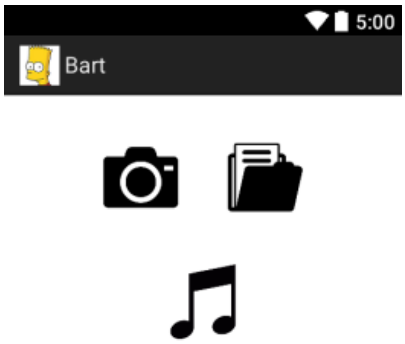


Figura 13 – Tela principal da aplicação.

3.3 Processamento de imagem

Como o foco principal da aplicação é para áudio-transcrever textos a partir de imagens, este é um dos principais módulos da aplicação, e tem como principal objetivo preparar a imagem para servir de entrada para o módulo de reconhecimento. Para realizar este trabalho foi implementado um sistema baseado no apresentado em (SáNCHEZ et al., 2011), e exemplificado na Figura 14, a escolha deste sistema ocorre por sua robustez, bem como pelo fato de sua implementação já ter sido executada no projeto de iniciação científica (GOUVEIA; BEZERRA, 2012).

Figura 14 – Pré processamento.

3.3.1 Binarização

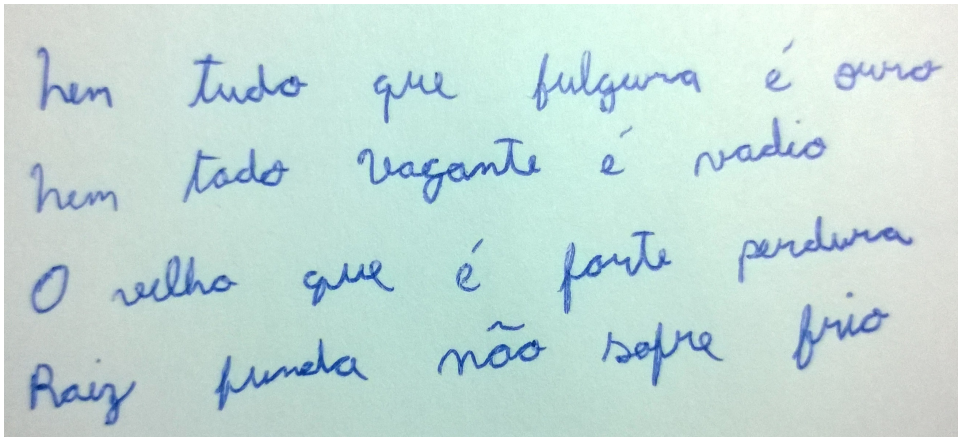
Como as imagens capturadas por câmeras de dispositivos móveis normalmente são coloridas, o primeiro passo a ser executado é a passagem da imagem para a escala de cinza (Figura 15), este passo é realizado aplicando a equação 3.1

$$p(x,y) = 0,299 \cdot r(x,y) + 0,587 \cdot g(x,y) + 0,114 \cdot b(x,y) \tag{3.1}$$

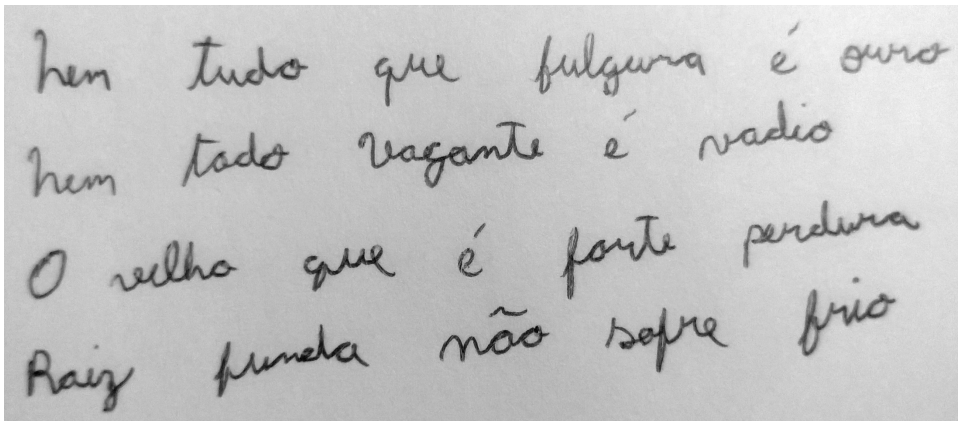
Onde:
$p(x,y)$ representa o pixel na posição (x,y) da imagem de saída,
r, g e $b(x,y)$ representam a intensidade dos canais vermelho, verde e azul do pixel (x,y) na imagem colorida respectivamente.

Com a imagem em tons de cinza é então aplicado um filtro de binarização para separar a região de texto da região de fundo. O método escolhido foi o apresentado por Gatos em (GATOS; PRATIKAKIS; PERANTONIS, 2006). Primeiro a imagem em tons de cinza é submetida a um filtro de suavização de Wiener (JAIN, 1989) de tamanho 3x3 gerando a imagem suavizada I, em seguida é realizada uma etapa de aproximação grosseira da região através de uma binarização de Sauvola (SAUVOLA; PIETIKÄINEN, 2000) 2.2 com $k = 0,2$, gerando a imagem que chamaremos de S. Em posse de S, o próximo passo é o de estimativa de background. Para esta etapa, pegamos a imagem original, após a aplicação da suavização, juntamente com S e avaliamos o valor de cada pixel, se o valor em S for 0, então mantemos o pixel da imagem suavizada, caso contrário, realizamos um processo de interpolação entre os vizinhos, gerando a imagem B, descrito em (**??**), onde d_x e d_y são o tamanho horizontal e vertical da janela de interpolação que deve ser de aproximadamente 2 vezes o tamanho de um caractere. Para gerar a imagem final binarizada é definido um limiar T, e é realizada uma subtração da imagem suavizada pela imagem B, e para cada pixel dessa nova imagem avalia-se se o este é menor que T, classificando-o como texto, ou maior que T, classificando-o como fundo. Um exemplo da binarização da Figura 15a é apresentado na Figura 15c.

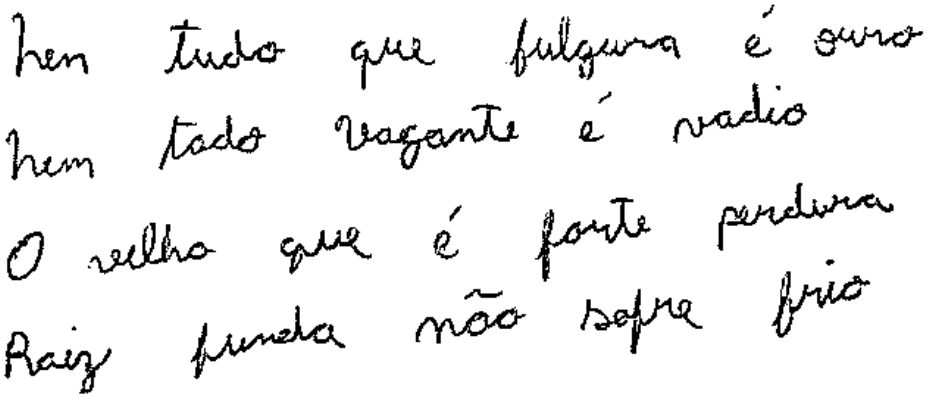
$$B(x,y) = \begin{cases} I(x,y) & seS(x,y) = 0 \\ \frac{\sum_{ix=x-d_x}^{x+d_x} \sum_{iy=y-d_y}^{y+d_y} (I(i_x,i_y) \cdot (1 - S(i_x,i_y)))}{\sum_{ix=x-d_x}^{x+d_x} \sum_{iy=y-d_y}^{y+d_y} (1 - S(i_x,i_y))} & seS(x,y) = 1 \end{cases} \tag{3.2}$$



(a)



(b)



(c)

Figura 15 – Conversão imagem colorida 15a para tons de cinza 15b, e posteriormente binarizada 15c.

3.3.2 Segmentação em palavras

O processo de segmentação envolve cinco etapas principais (Figura 16), suavização, esqueletização, detecção, agrupamento e finalmente a etapa de dilatação .

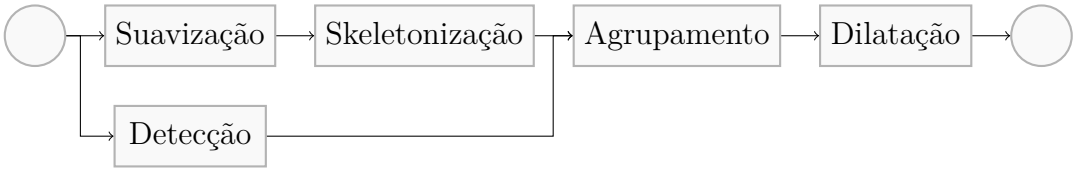


Figura 16 – Segmentação.

O objetivo da etapa de suavização é a obtenção das linhas da imagem, para isto aplica-se uma janela deslizante em cada pixel, onde se calcula o somatório do valor total das transições de fundo para objeto, o tamanho da janela é de um terço da largura da imagem original, com altura de 1 pixel.

A etapa de skeletonização serve para termos um ajuste fino na camada de suavização. Primeiro é aplicado uma binarização global na imagem suavizada usando-se Otsu. Em seguida, usando-se o algoritmo de Suen (ZHANG; SUEN, 1984) sobre essa imagem, temos a versão mínima das linhas.

Na detecção é feita o passo de separar os componentes conectados da imagem. Esta etapa segue o Algoritmo 1.

Algoritmo 1: Detecção de componentes conexos.

```
1  M ← matriz com as mesmas dimensões da imagem;
2  C ← 1;
3  pilha ← pilha vazia;
4  for i ← 0 to altura da imagem do
5      for j ← 0 to largura da imagem do
6          if imagem na posição (i, j) for objeto then
7              Adicionar a posição (i, j) na pilha;
8              while pilha não vazia do
9                  topo ← captura o topo da pilha;
10                 marca em m a posição topo com o valor do contador;
11                 adiciona todos os vizinhos em uma vizinhança 8 de topo que não
                   estejam marcados em m e que sejam objeto;
12             c ← c + 1;
```

Em posse das imagens geradas pelas etapas de skeletonização e de detecção é realizada a etapa de agrupamento, que consiste em: para cada segmento detectado, verificar a qual linha este pertence. Após ter sido detectada a qual linha cada componente pertence,

é realizada uma etapa de dilatação neste para melhor detectar palavras, dado que nem sempre as pessoas escrevem uma palavra de forma 100% contínua. Ao término da etapa de dilatação, temos a região de cada possível palavra encontrada na imagem, que servirá de entrada para a próxima fase do sistema. Um processo de segmentação está demonstrado na Figura 17.

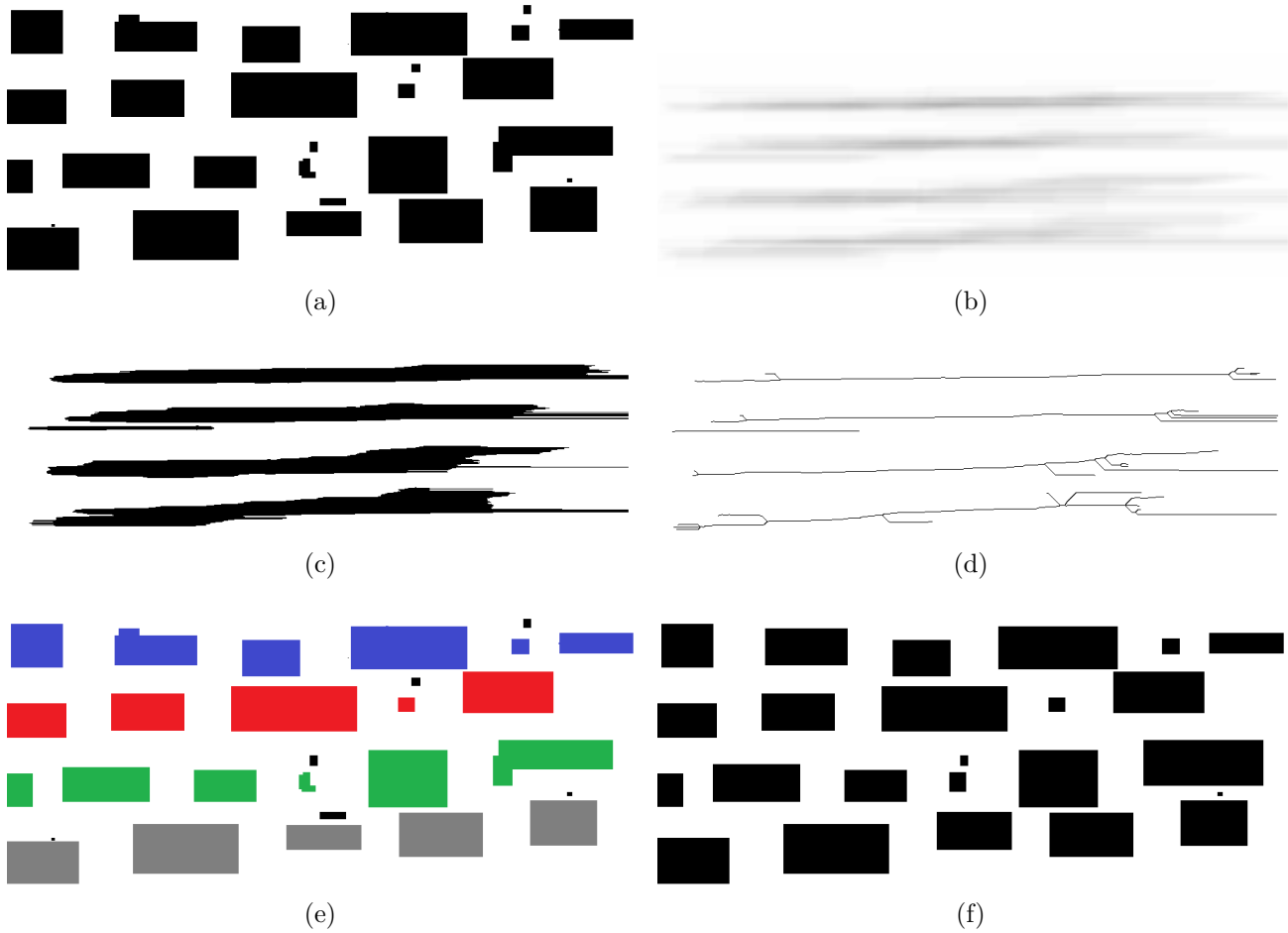


Figura 17 – Processo de segmentação da Figura 15c, começando pelas etapas de detecção de componentes 17a e suavização 17b e sua binarização 17c, seguido pela etapa de skeletonização 17d, detecção de componentes por linha 17e e dilatação 17f.

3.4 Classificação

O sistema de classificação foi baseado no proposto por Graves em (GRAVES; SCHMIDHUBER, 2009), que tem se mostrado como um dos principais e melhores sistemas para o reconhecimento de escrita cursiva offline, tendo resultados bastante relevantes nos últimos anos como apresentado por Louradour (LOURADOUR; KERMORVANT, 2014), Doetsch (DOETSCH; KOZIELSKI; NEY, 2014) e Bezerra (BEZERRA; ZANCHETTIN; ANDRADE, 2012). A arquitetura proposta por Graves de uma rede neural recorrente multidimensional segue o modelo apresentado na Figura 18, onde cada neurônio da rede

recebe, fora os estímulos da camada anterior como uma rede neural MLP convencional, estímulos dos seus antecessores diretos. Esta característica faz com que o neurônio tenha o conhecimento local da informação que está processando.

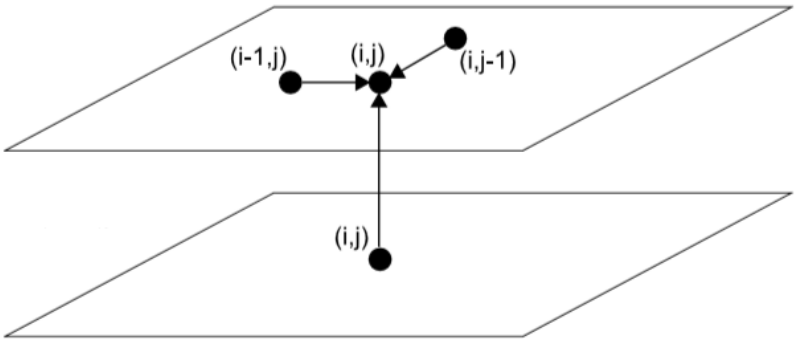


Figura 18 – Modelo multidimensional recorrente (GRAVES; FERNANDEZ; SCHMIDHUBER, 2007).

Como fica evidente na Figura 18, a ordem em que a informação de cada camada é lida influencia diretamente o comportamento da rede, por exemplo, em uma imagem podemos ler a informação a partir dos quatro cantos desta, e ter resultados diferentes, levando em consideração isto. Admitindo que a leitura da informação em cada uma das direções da imagem é importante, Graves faz uma expansão da rede, onde cada camada da rede gera 2^n versões da camada escondida, uma para cada direção que pode ser percorrida a informação. Após cada *layer* ser processado, uma camada de *collapse*, realiza uma junção das camadas anteriores. Cada *layer* da rede é composto por células LSTM. A camada de saída da rede é composta por uma camada chamada de *connectionist temporal classification* (CTC) (GRAVES et al.,), que estende o poder das redes neurais de classificarem entradas para padrões já conhecidos (cada entrada tem como saída um padrão), para poderem reconhecer sequências de padrões através de uma análise de probabilidades das saídas da rede. Mais detalhes sobre esta arquitetura podem ser encontrados em (GRAVES et al., 2012).

3.5 Síntese

O processo de síntese de áudio foi desenvolvido em colaboração com o Framework FIVE (MACIEL; CARVALHO, 2010). Como a complexidade desta é alta e sua portabilidade para um dispositivo móvel exigiria um grande esforço, optou-se por desenvolver um serviço web simples. O Framework FIVE utiliza de um conjunto de características próprios juntamente com um modelo escondido de Markov para gerar a síntese do seu sistema.

Este serviço tem por objetivo receber um texto e retornar para o usuário a versão em áudio deste. A parte web deste serviço foi desenvolvido utilizando a api jersey (JERSEY,

2015) para java, que simplifica o desenvolvimento de serviços web baseados em REST (IBM, 2015) para uma chamada de método em java.

4 Resultados

Neste capítulo são apresentados e discutidos os resultados encontrados com o desenvolvimento do sistema.

4.1 Base de reconhecimento em português

Um dos desafios encontrados para o desenvolvimento do sistema foi o de encontrar uma base de palavras em português. Como não foi encontrada nenhuma base de palavras em português, decidimos criar uma própria a partir de uma parceria com a Stefanini Document Solutions (SOLUTIONS, 2015) (DS). A DS disponibilizou cerca de 1000 imagens de documentos manuscritos nacionais, possibilitando a construção de uma base com cerca de 10000 palavras manuscritas. O trabalho de segmentação e indexação da base de palavras contou com a colaboração dos alunos de iniciação científica Airton, Italo e Gabriel, orientados pelo Prof. Byron. A base de palavras foi batizada como RPPDI-DS-WBR1 e serve aos propósitos de pesquisa do grupo RPPDI e da DS. Como a base possui uma baixa diversidade do conjunto de palavras, foi decidido que o sistema seria treinado por dicionário. Nesse modelo de treinamento ao invés da rede ser treinada para acertar a palavra em si, ou seja, a rede ter $n + 1$ saídas onde cada saída corresponde a uma palavra e a última/primeira corresponde a *blank*, as palavras são quebradas em trechos menores, no nosso caso em letras, e a rede é treinada para acertar a sequência correspondente a esta palavra.

Antes do treinamento foram realizadas uma série de operações sobre as imagens para melhorá-las para entrada do modelo, como correção de inclinação da palavra, correção de alinhamento das letras, normalização da altura das imagens e um processo de binarização.

Para o treinamento do sistema, a base foi dividida em três partes: um conjunto de treinamento com 70% das imagens, um conjunto de validação com 15%, e um conjunto de teste com os 15% restantes. A arquitetura da rede foi inspirada na proposta em (MOYSSET et al., 2014), contendo 3 camadas escondidas, bloco de entrada de tamanho 2×2 e blocos intermediários de tamanho 2×4. O treinamento foi realizado utilizando-se o método de gradiente descendente com momento de 0.9 e taxa de aprendizagem de 0.0001. Embora esta arquitetura tenha demonstrado bons resultados como o apresentado em (GRAVES et al., 2009), devido a baixa diversidade da base os resultados se mostraram medianos, tendo alcançado uma taxa de erro de 30% para o reconhecimento de cada rótulo (letra), e uma taxa geral de erro de 60% para as sequências. Ou seja, a rede consegue de certa forma identificar os principais caracteres da base, mas não consegue gerar informação suficiente para conseguir reconhecer uma palavra por completo. Para efeito de comparação uma das

bases de referência na área de reconhecimento de escrita, a IAMDB (MARTI; BUNKE, 2002) possui 115.320 imagens de palavras individuais.

4.2 Execução do sistema

Dado que os resultados com a base em português não obtiveram um resultado bastante satisfatório, optou-se por utiliza uma base já treinada para reconhecimento de nomes de meses em português, emprestada também pela DS, esta base tem uma taxa de acerto de 77,25%. Para os testes foram tiradas fotos da escrita em um quadro branco, utilizando o aplicativo desenvolvido, os testes foram realizados utilizando-se um tablet Galaxy Tab 2, tirando fotos com resolução de 800x600. No geral o sistema foi capaz de reconhecer corretamente e em um tempo aceitável, cerca de 8 segundos por imagem para as etapas de processamento e classificação, e 1 segundo por palavra para síntese, ou 2 segundos para o caso do texto completo, cada palavra presente no texto. Com o intuito de deixar o sistema com melhor usabilidade para usuários portadores de deficiência visual, a interface toda do sistema se encontra em uma única tela na qual o usuário tem a opção de tirar uma foto, escolher um arquivo no sistema, ou reproduzir o último áudio processado, vide Figura 19

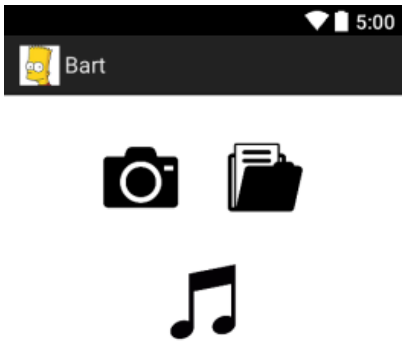


Figura 19 – Tela principal da aplicação.

5 Considerações Finais

O sistema mostrou ter um bom desempenho ao ser executado em dispositivos móveis, sendo capaz de rodar os módulos de captura, processamento de imagem e classificação no próprio dispositivo de forma satisfatória, e também sendo capaz de executar o módulo de síntese na web e reproduzir o resultado no módulo de saída de forma satisfatória.

Como resultado maior do desenvolvimento deste trabalho podemos citar a publicação do artigo *Handwriting recognition system for mobile accessibility to the visually impaired people* na conferência *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* de 2014 (GOUVEIA et al.,).

5.1 Trabalhos futuros

Embora o sistema tenha se provado como uma possibilidade real, algumas melhorias podem ser executadas neste, com o intuito da busca de um melhor resultado na sua execução. A primeira vista, por se tratar de um sistema voltado para dispositivos móveis com foco em imagens obtidas a partir da câmera, é necessário que seja realizada uma etapa inicial no processamento da imagem que seria responsável por corrigir de forma dinâmica a perspectiva da imagem. Outro ponto a ser melhorado na parte de processamento de imagem é o de realizar a correção de inclinação na etapa de detecção de linhas.

Por parte do sistema de classificação seria interessante conseguir incrementar a base de palavras com o intuito de se conseguir melhorar o modelo para termos um produto com melhor utilidade para o público.

Conhecendo a realidade brasileira e de países mais pobres, e com menor acesso a internet, é de total interesse fazer com que o sistema de síntese rode diretamente no dispositivo, sem a necessidade da chamada de um sistema web.

Referências

ANDROID. *Android Developer*. 2015. Disponível em: <http://developer.android.com/index.html>.

BEZERRA, B. L. D.; ZANCHETTIN, C.; ANDRADE, V. B. de. A mdrnn-svm hybrid model for cursive offline handwriting recognition. In: *Artificial Neural Networks and Machine Learning–ICANN 2012*. [S.l.]: Springer, 2012. p. 246–254.

BUNKE, H. Recognition of cursive roman handwriting: past, present and future. In: IEEE. *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*. [S.l.], 2003. p. 448–459.

BUNKE, H.; BENGIO, S.; VINCIARELLI, A. Offline recognition of unconstrained handwritten texts using hmms and statistical language models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 26, n. 6, p. 709–720, 2004.

DOETSCH, P.; KOZIELSKI, M.; NEY, H. Fast and robust training of recurrent neural networks for offline handwriting recognition. In: IEEE. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. [S.l.], 2014. p. 279–284.

DUTOIT, T. High-quality text-to-speech synthesis: An overview. *Journal Of Electrical And Electronics Engineering Australia*, IREE INSTITUTION OF RADIO AND ELECTRONICS, v. 17, p. 25–36, 1997.

EZAKI, N. et al. Improved text-detection methods for a camera-based text reading system for blind persons. In: IEEE. *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. [S.l.], 2005. p. 257–261.

FERREIRA, S.; GARIN, V.; GOSSELIN, B. A text detection technique applied in the framework of a mobile camera-based application. In: *Proceedings of the First International Workshop on Camera-based Document Analysis and Recognition (CBDAR)*. [S.l.: s.n.], 2005.

GATOS, B.; PRATIKAKIS, I.; PERANTONIS, S. J. Adaptive degraded document image binarization. *Pattern recognition*, Elsevier, v. 39, n. 3, p. 317–327, 2006.

GERS, F. Long short-term memory in recurrent neural networks. *Unpublished PhD dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*, Citeseer, 2001.

GERS, F. A.; SCHRAUDOLPH, N. N.; SCHMIDHUBER, J. Learning precise timing with lstm recurrent networks. *The Journal of Machine Learning Research*, JMLR. org, v. 3, p. 115–143, 2003.

GONZALEZ, R. C. *Digital image processing*. [S.l.]: Pearson Education.

GOUVEIA, F. M.; BEZERRA, B. L. D. Segmentacao de texto cursivo em dispositivos moveis. *Encontro de Pos Graduacao Pesquisa e Extensao 2012 da UPE.*, UPE, 2012.

GOUVEIA, F. M. et al. Handwriting recognition system for mobile accessibility to the visually impaired people. In: *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. [S.l.: s.n.].

GRAVES, A. et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: ACM. *Proceedings of the 23rd international conference on Machine learning*. [S.l.]. p. 369–376.

GRAVES, A.; FERNANDEZ, S.; SCHMIDHUBER, J. Multi-dimensional recurrent neural networks. In: . [S.l.: s.n.], 2007.

GRAVES, A. et al. A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 31, n. 5, p. 855–868, 2009.

GRAVES, A. et al. *Supervised sequence labelling with recurrent neural networks*. [S.l.]: Springer, 2012. v. 385.

GRAVES, A.; SCHMIDHUBER, J. Offline handwriting recognition with multidimensional recurrent neural networks. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2009. p. 545–552.

HOCHREITER, S. et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. A field guide to dynamical recurrent neural networks. IEEE Press, 2001.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

IBM. *RESTful Web services: The basics*. 2015. Disponível em: <http://www.ibm.com/developerworks/library/ws-restful/>.

IDC. *Smartphone OS Market Share, Q1 2015*. 2015. Disponível em: <http://www.idc.com/prodserv/smartphone-os-market-share.jsp>.

ITU. *ICT Facts and Figures*. 2015. Disponível em: <http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf>.

JAIN, A. K. *Fundamentals of digital image processing*. [S.l.]: Prentice-Hall, Inc., 1989.

JERSEY. *Jersey*. 2015. Disponível em: <https://jersey.java.net/>.

LECHETA, R. R. *Google Android-3ª Edição: Aprenda a criar aplicações para dispositivos móveis com o Android SDK*. [S.l.]: Novatec Editora, 2013.

LEMMETTY, S. Review of speech synthesis technology. *Helsinki University of Technology*, 1999.

LIANG, J.; DOERMANN, D.; LI, H. Camera-based analysis of text and documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, Springer, v. 7, n. 2-3, p. 84–104, 2005.

LOURADOUR, J.; KERMORVANT, C. Curriculum learning for handwritten text line recognition. In: IEEE. *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. [S.l.], 2014. p. 56–60.

MACIEL, A.; CARVALHO, E. Five–framework for an integrated voice environment. In: *Proceedings of International Conference on Systems, Signal and Image Processing, Rio de Janeiro*. [S.l.: s.n.], 2010.

MARTI, U.-V.; BUNKE, H. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, Springer, v. 5, n. 1, p. 39–46, 2002.

MOYSSET, B. et al. The a2ia multi-lingual text recognition system at the second maurdor evaluation. In: IEEE. *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. [S.l.], 2014. p. 297–302.

NIBLACK, W. *An Introduction to Digital Image Processing*. Birkeroed, Denmark, Denmark: Strandberg Publishing Company, 1985. ISBN 87-872-0055-4.

O'SHAUGHNESSY, D. *Speech communication: human and machine*. [S.l.]: Universities press, 1987.

OTSU, N. A threshold selection method from gray-level histograms. *Automatica*, v. 11, n. 285-296, p. 23–27, 1975.

PLÖTZ, T.; THURAU, C.; FINK, G. A. Camera-based whiteboard reading: New approaches to a challenging task. In: CITESEER. *International Conference on Frontiers in Handwriting Recognition*. [S.l.], 2008. p. 385–390.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.

SAMSUNG. *Samsung Galaxy Note*. 2015. Disponível em: <http://www.samsung.com/global/microsite/galaxynote/note/spec.html?type=find>.

SáNCHEZ, A. et al. Automatic line and word segmentation applied to densely line-skewed historical handwritten document images. *Integr. Comput.-Aided Eng.*, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 18, n. 2, p. 125–142, abr. 2011. ISSN 1069-2509. Disponível em: <http://dl.acm.org/citation.cfm?id=1971744.1971747>.

SAUVOLA, J.; PIETIKÄINEN, M. Adaptive document image binarization. *Pattern recognition*, Elsevier, v. 33, n. 2, p. 225–236, 2000.

SOLUTIONS, D. *Stefanini Document Solutions*. 2015. Disponível em: <http://documentsolutions.com.br/>.

STAMATOPOULOS, N. et al. Icdar 2013 handwriting segmentation contest. In: IEEE. *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. [S.l.], 2013. p. 1402–1406.

WHO. *World report on disability*. 2015. Disponível em: <http://www.who.int/disabilities/world_report/2011/en/>.

ZHANG, G. P. Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, IEEE, v. 30, n. 4, p. 451–462, 2000.

ZHANG, T.; SUEN, C. Y. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, ACM, v. 27, n. 3, p. 236–239, 1984.