Aprendizagem dos Números para Crianças com o Uso do Kinect

Kleber M. Mesquita[1], Robson S. Siqueira[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) – Maracanaú, CE – Brasil
{kleber099,siqueira.robson.dasilva}@gmail.com

Abstract. *Development of new technologies to human-computer interaction has receive the most interest of researchers in developing new techniques and applications that facilitate and qualify people's lives. The popularization of depth sensors such as Microsoft Kinect has enabled progress in techniques that had many restrictions on the use of RGB images, minimizing occlusion problems like skin recognition. This paper presents a technique to identify numbers with hand gestures applied to interactive teaching children. The presented method aims to assist professionals in early childhood education to design programs that can get information about the learning progress.*

Resumo. *Com o rápido surgimento de novas tecnologias, a interação homem-máquina tem renovado o interesse de pesquisadores no desenvolvimento de novas técnicas e aplicações que facilitem e qualifiquem a vida das pessoas. A popularização de sensores de profundidade como o Microsoft Kinect, tem permitido evoluir em técnicas que possuíam muitas restrições com o uso de imagens RGB, minimizando problemas como oclusão e reconhecimento de padrões de pele. Este artigo apresenta uma técnica para identificar números com os gestos das mãos, aplicado ao ensino interativo de crianças. O método apresentado tem por objetivo auxiliar os profissionais em educação infantil a projetar programas que possam obter informações sobre o andamento do aprendizado.*

1. Introdução

Reconhecimento de gestos das mãos é de grande importância para a interação homem-máquina (HCI), por causa de suas inúmeras aplicações na realidade virtual, reconhecimento de linguagem de sinais e jogos de computador [2]. Apesar de muitos trabalhos anteriores, métodos de reconhecimento de gestos de mão tradicionais, com base na visão [7, 8] ainda estão longe de ser satisfatórios para aplicações na vida real. Devido às limitações dos sensores ópticos, a qualidade das imagens capturadas é sensível às condições de iluminação e de fundo ofuscado, assim não é capaz de detectar e identificar as mãos de forma robusta, o que afeta em grande medida o desempenho do reconhecimento de gestos [1].

Graças ao recente desenvolvimento de câmeras de profundidade de baixo custo, por exemplo, o Kinect, novas oportunidades para o reconhecimento de gesto surgem, esse trabalho propõe reconhecer gestos das mãos utilizando o Kinect, verificando quantos dedos estão sendo mostrados nas cenas capturadas, possibilitando realizar a quantificação, identificação de formas e atribuir posteriormente a execução de ações ao reconhecer cada padrão estabelecido. Tornando possível gerar indicadores, por exemplo, sobre o processo de aprendizagem de números para crianças que estão no nível adequado esse tipo de conhecimento, bem como para crianças ou adultos que

estejam reaprendendo, após processos traumáticos de perda total ou parcial de memória ou de outras faculdades cognitivas.

A geração desses indicadores, sob a coordenação de um profissional da área, pode ser obtida com testes não supervisionados aos usuários finais, de maneira a auxiliá-los na avaliação do nível de cognição visual, auditiva, motora ou tempo de reação, dentre outras informações que possam ser disponibilizadas e sejam importantes para os especialistas.

Todo esse trabalho será realizado com processamento e análise das imagens de profundidade do Kinect com o intuito de identificar a quantidade de dedos dos gestos das mãos e apresentar um modelo para aplicações no aprendizado infantil.

2. O Kinect e sua utilização na Educação Infantil

O processo de aprendizagem vai além do ambiente da sala de aula e está presente em todos os momentos em que uma troca de experiência, não necessariamente entre professor e aluno, possa vir a acrescentar conhecimento ao indivíduo [3]. Devem-se criar diferentes formas de organização da classe, dos tempos e espaços didáticos, dos objetos, recursos e estratégias pedagógicas [5]. Diferenciar é organizar as interações e as atividades, de modo que cada aluno seja constantemente confrontado com as situações mais fecundas para ele, que sejam do seu interesse ou que seja um obstáculo à construção do conhecimento. Assim, o ensino diferenciado implica a utilização de diversas estratégias didáticas [4].

Através da NUI *(Natural User Interface*) pode se criar um ambiente de forma mais natural, por que se baseia em um ambiente feito por interações. Muitas dessas interações são rotineiras e conhecidas por usuários, no caso crianças. A interface que proporciona a NUI é um *design* de aplicação que permite explorar do usuário a melhor forma de interagir com o objeto em questão, fazendo com que a experiência de uso do produto se aproxime do mais confortável, intuitivo e simples possível, para o usuário que está utilizando a aplicação [6].

Em pouco tempo Kinect pouco a ser utilizado como objeto de estudo e desenvolvimento de aplicações em Realidade Aumentada (RA) devido a sua capacidade de reconhecimento de movimento, comando de voz, sem a necessidade de algum outro dispositivo adicional. Dessa forma, permite pessoas interagirem com os jogos, sistemas e dispositivos, utilizando se do movimento do próprio corpo [3]. Além disso, é capaz de perceber a terceira dimensão (profundidade).

Uma das grandes vantagens de utilizar imagens de profundidade no lugar de imagens RGB são as restrições ligadas, principalmente, à iluminação do ambiente. A detecção de pele, utilizada para detectar as mãos, é muito dependente da iluminação que pode alterar os parâmetros de reconhecimento pré-definidos pelo sistema [9]. O sensor de profundidade não precisa de luz para funcionar e por isso não sofre com problemas de sombreamento se houver uma fonte luminosa predominante, de forma similar à imagem em tons de cinza, tem um canal que possui uma janela com largura e altura características. A diferença está no significado do valor do pixel, no caso da imagem em tons de cinza, ela equivale ao tom da cor cinza; já no caso da imagem de profundidade, diz respeito à distância entre o sensor e a superfície do objeto detectado. Outra diferença importante é que a imagem em tons de cinza possui 8 bits, correspondendo à 256 tons; no caso da imagem de profundidade são 11 bits, dos quais 10 correspondem à distância

da superfície do objeto ao sensor e o outro indica se o ponto é válido ou não. No caso, a distância pode variar de 0 a 1023, caso contrário, o pixel é desconsiderado.

3. Reconhecimento de Gestos da Mão

3.1. Ambiente de Pesquisas e Testes

O ambiente de desenvolvimento e testes desse trabalho foi realizado em sistema operacional Linux, distribuição Debian Wheezy. Os algoritmos na pesquisa foram inscritos nas linguagens C/C++, utilizando a biblioteca OpenCv, afim de manipular, tratar e processar as imagens. Diversos usuários realizaram gestos com as mãos na frente do Kinect e as imagens geradas foram utilizadas pelos algoritmos desta pesquisa a fim de reconhecer e obter informações importantes a cerca das mãos. (O uso da aplicação deverá ser utilizado em computadores conectados com o Kinect.)

3.2. Método de Reconhecimento da Mão

A Fig. 3.1 mostra o fluxograma das etapas que foram necessárias no processo de reconhecimento da mão, que foi composto da aquisição da imagem pelo o kinect até a delimitação das regiões que são dedos e extração das informações importantes a cerca da mão.

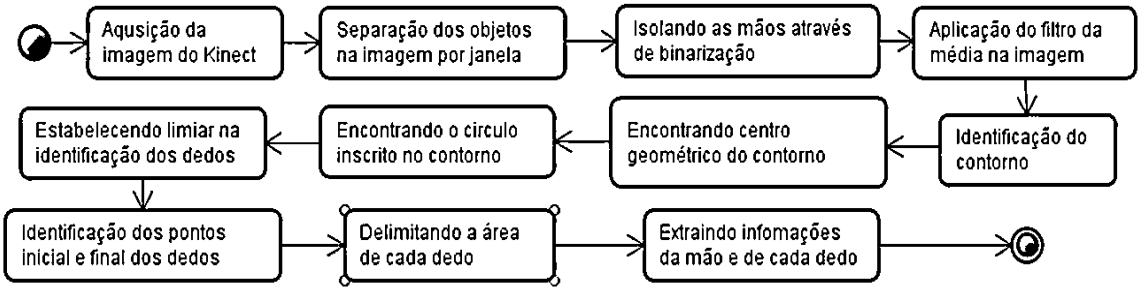


Figura 3.1. Fluxograma de reconhecimento da mão

Como foi mencionado na seção 2, o sensor de profundidade do Kinect atribui valores conforme a distância dos objetos na cena. Para capturar somente as mãos, foi utilizada uma janela com intervalo entre 300 e 500 na coordenada de profundidade, isto equivale à uma distância entre 100 e 150 cm do sensor, que pode ser modificada de acordo com o ambiente onde será realizado o teste. Foi construído um algoritmo que atribui para atribuir cores aos objetos identificados de acordo com a distância que e encontra os objetos estão do sensor

A próxima etapa constitui em binarizar a imagem capturada do Kinect, para que somente dois níveis de cores fossem utilizados. Caso o objeto estivesse no intervalo de janela estabelecido, seria atribuída à cor branca (valor 255), se não estivesse seria atribuída à cor preta (valor 0). Com isso, foi possível separar as mãos do restante dos outros objetos que estavam presentes na cena. A Fig. 3.2 apresenta a imagem após o processo de binarização.

Figura 3.2. Imagem após a aplicação do algoritmo de divisão por janelas e binarização da imagem

Em seguida foi aplicado o filtro da média na imagem utilizando o tamanho 5x5. Com isso foi possível suavizar as imagens adquiridas e retirar os diversos ruídos que o sensor do Kinect detectou. A imagem tratada com o filtro da média pode ser visualizada na Fig. 3.3.a.

(a) (b) (c)

Figura 3.3. (a): Comparação entre a imagem binarizada e aplicação do filtro da média. (b): Identificação do contorno da mão. (c): Centro geométrico do contorno

Após o tratamento com o filtro da média, foi possível identificar o contorno das mãos, necessário para obter informações que ajudarão a demarcar a região onde se encontram os dedos na imagem e a quantidade destes que está sendo mostrada em cada mão. O contorno de uma mão pode ser visualizado na Fig. 3.3.b.

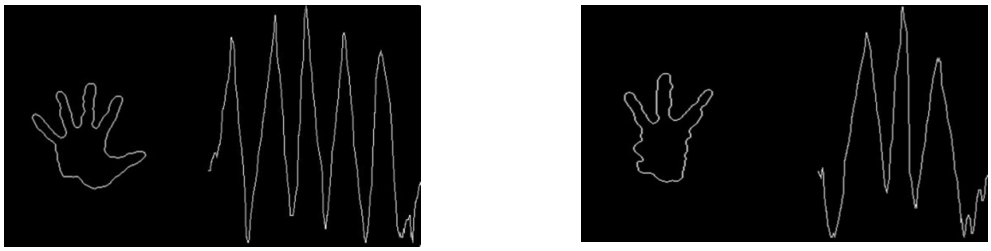
Com o contorno de cada mão encontrado, a próxima etapa foi encontrar a centroide (ou centro geométrico) de cada um. Centroide é o ponto central de uma forma geométrica qualquer, caso a forma geométrica seja homogênea, então o centroide coincide com o centro de massa. Nos casos em que essa forma não é homogéneo, então esse ponto coincide com o centro gravitacional. O Centro geométrico do contorno da mão pode ser visualizado na Fig. 3.3.c.

Foi utilizado o ponto central da mão para encontrar a distância entre cada ponto do contorno em relação ao seu centro geométrico. Para calcular essa distância, foi usado distância Euclidiana entre dois pontos, dada por:

$$D = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$

O ponto $P = (p_x, p_y)$ representa cada ponto que faz parte do contorno e $Q = (q_x, q_y)$ é o ponto central que está dentro do contorno. A partir desses dois pontos, é possível calcular a distância D.

Em seguida foram construídos gráficos para representar as formas adquiridas pelo Kinect. A Fig. 3.4.a apresenta o gráfico de uma mão, onde é representada a distância entre o centro geométrico em relação a cada ponto do contorno. Os pontos do contorno mais distantes do centro geométrico estão concentrados nos picos do gráfico, e os pontos com menor distância em relação ao centro estão nas regiões mais inferiores no gráfico. Os picos são as pontas de cada dedo. A imagem 3.4.a contém cinco picos referentes aos cinco dedos da mão esquerda, caso seja detectado menos dedos, menos picos serão apresentados, conforme Fig. 3.4.b.



(a) (b)

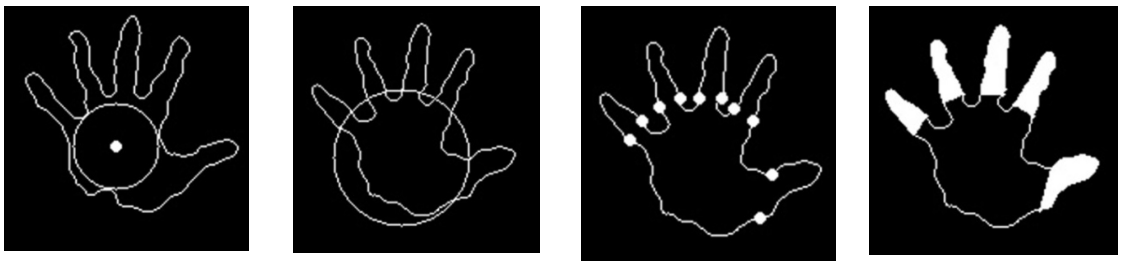
Figura 3.4. (a): Gráfico da distância entre cada ponto do contorno e o centro da mão. (b): Gráfico da mão com três dedos

Para identificar a quantidade de dedos, definiu-se um limiar L que identifique a presença de dedos nas imagens adquiridas. Limiar é um valor limite que divide duas ou mais classes. Para esse trabalho, objetivou-se encontrar um L que dividisse as regiões que contivesse dedos daquelas que não contivesse. Um aspecto importante de definição deste limiar é que ele deve servir para tamanhos de mãos diferentes, pois o algoritmo pode ser aplicado a crianças, jovens ou adultos. Isso somente é possível se for utilizado um valor adimensional, com base em alguma característica da mão. Neste caso, foi utilizado como parâmetro o raio inscrito na mão, como sugerido por Zou-Zang[1].

Para encontrar o círculo inscrito dentro do contorno, foi calculada a menor distância entre o centroide e o contorno de cada mão. Essa distância calculada é o raio R do círculo inscrito de cada contorno, o qual pode ser visualizado na Fig. 3.5.a. Com o raio do círculo inscrito ao contorno, utilizou-se o valor de R, a fim de estabelecer um L que vai identificar a presença de dedos nos contornos obtidos. Ao utilizar uma constante c qualquer, estabeleceu-se uma fórmula para encontrar um valor para L, dada por $L = c \times R$. Dentro do universo de 1000 imagens adquiridas para teste de 10 usuários diferentes, o valor de $1{,}7$ para c mostrou se o mais adequado. Pela Fig. 3.5.b, percebe-se que ao utilizar esse limiar, o círculo da figura intercepta todos os dedos do contorno, então, as distâncias $D \geq L$ representam as regiões que são dedos no contorno encontrado.

Em seguida foi delimitada a sequência de pontos de cada dedo, detectando o limite inicial e o final de cada um deles, assim foram obtidas informações importantes como, por exemplo, a área, quantidade de pontos em cada dedo e estabelecidos critérios para classificar se as regiões encontradas eram realmente dedos. A Fig. 3.5.c apresenta os pontos que fazem intercessão entre o contorno e o círculo de raio L, os quais delimitam o início e fim de cada dedo.

Ao obter os limites de cada dedo e juntos com seus pontos de contorno, calculou-se a área pelo número de pixels existentes dentro do contorno. Na Fig. 3.5.d pode-se visualizar as regiões da imagem que foram detectadas como dedos denotados pelas áreas pintadas pela cor branca. São justamente os pontos cujo $D \geq L$, caso $D < L$, tais pontos não pertencem a dedos e suas regiões não serão identificados como tal.



(a) (b) (c) (d)

Figura 3.5. (a): Circulo inscrito no contorno de raio R. (b): Circulo com $L = 1{,}7 \times R$. (c): Pontos que delimitam cada dedo. (d): Detecção de dedos

4. Testes

Como visto na seção 3.2, é possível obter informações importantes a cerca da mão. Para os testes, foi utilizado a área da mão A, as áreas de cada dedo: polegar A_p, indicador A_i, médio A_m, anelar A_a e mindinho A_{mi}; a quantidade de pontos de cada dedo: polegar Q_p, indicador Q_i, médio Q_m, anelar Q_a e mindinho Q_{mi}; a distância da ponta de cada dedo em relação ao centro: polegar D_p, indicador D_i, médio D_m, anelar D_a e mindinho D_{mi} e a excentricidade de cada dedo: polegar E_p, indicador E_i, médio E_m, anelar E_a e mindinho E_{mi}.

Foram utilizados os valores acima para calcular dados estatísticos das características das mãos, à média $\overline{x}$ e o desvio padrão σ, para estabelecer um intervalo de confiança. A média das áreas de cada dedo em relação à área da mão: correspondeu à $\overline{(A_p/A)}$, $\overline{(A_i/A)}$, $\overline{(A_m/A)}$, $\overline{(A_a/A)}$ e $\overline{(A_{mi}/A)}$; a média da quantidade de pontos de cada mão: $\overline{Q_p}$, $\overline{Q_i}$, $\overline{Q_m}$, $\overline{Q_a}$ e $\overline{Q_{mi}}$; a média da distância da ponta de cada dedo ao centro: $\overline{D_p}$, $\overline{D_i}$, $\overline{D_m}$, $\overline{D_a}$ e $\overline{D_{mi}}$ e a média da excentricidade de cada dedo: $\overline{E_p}$, $\overline{E_i}$, $\overline{E_m}$, $\overline{E_a}$ e $\overline{E_{mi}}$.

Sejam x todos os valores obtidos das formas ou regiões capturadas, se $x \in (\overline{x} - \sigma, \overline{x} + \sigma)$, x é uma característica de uma mão, então, a forma ou região é de uma mão. Caso os valores não estejam no intervalo de confiança, as formas ou regiões não podem ser consideradas como mão.

5. Resultados

A Tabela 5.1 mostra os valores calculados das estatísticas das imagens adquiridas de uma base de dados, com amostras de 1000 imagens que tenha somente mãos.

Tabela 5.1 Valores estatísticos das características da mão

Polegar							
$\overline{(A_p/A)}$	σ	$\overline{Q_p}$	σ	$\overline{D_p}$	σ	$\overline{E_p}$	σ
0.041149	0.012238	29.619835	6.187760	71.475207	3.160326	0.239564	0.101158
Indicador							
$\overline{(A_i/A)}$	σ	$\overline{Q_i}$	σ	$\overline{D_i}$	σ	$\overline{E_i}$	σ
0.061531	0.006850	43.057851	6.456814	83.739669	5.246976	0.227218	0.060416
Médio							
$\overline{(A_m/A)}$	σ	$\overline{Q_m}$	σ	$\overline{D_m}$	σ	$\overline{E_m}$	σ
0.071748	0.009371	37.060606	6.830291	88.703857	5.918449	0.233051	0.061386
Anelar							
$\overline{(A_a/A)}$	σ	$\overline{Q_a}$	σ	$\overline{D_a}$	σ	$\overline{E_a}$	σ
0.051712	0.009089	29.049587	6.690004	80.106061	6.221302	0.279684	0.080880
Mindinho							
$\overline{(A_{mi}/A)}$	σ	$\overline{Q_{mi}}$	σ	$\overline{D_{mi}}$	σ	$\overline{E_{mi}}$	σ
0.028489	0.009147	24.588154	6.079458	70.998623	5.891783	0.276541	0.113298

Nos diversos testes realizados em 95% dos casos os valores estavam dentro do intervalo de confiança, os demais casos não foram classificados como mãos,

pois eles estavam fora do intervalo de confiança. A Tabela 5.2 mostra alguns valores obtidos nos testes realizados.

Tabela 5.2 Valores das características da mão

Polegar				Indicador				Médio			
A_p/A	Q_p	D_p	E_p	A_i/A	Q_i	D_i	E_i	A_m/A	Q_m	D_m	E_m
0.03065	25	69	0.161692	0.059473	49	88	0.249996	0.062905	34	91	0.213605
0.030820	24	72	0.324245	0.058002	44	87	0.235579	0.064864	41	92	0.220216
0.044286	29	72	0.174301	0.060145	46	84	0.183015	0.064169	33	91	0.254420
Anelar				Mindinho							
A_a/A	Q_a	D_a	E_a	A_{mi}/A	Q_{mi}	D_{mi}	E_m				
0.044303	30	75	0.216659	0.030071	30	65	0.286135				
0.042916	29	86	0.261702	0.022781	27	66	0.262790				
0.046045	28	83	0.230703	0.030644	19	71	0.279992				

Observa-se pela tabela que cada dedo tem características próprias e esse valores estatísticos pode ser utilizados para identificar um dedos particular, pois a faixa de valores de cada característica está em intervalo diferente de valores para cada dedo.

6. Aplicação de Aprendizagem de Números

Conforme o usuário vai realizando gestos em frente do sensor do Kinect o aparelho vai capturando seus movimentos. É importante que a mão do usuário esteja em destaque na frente do sensor. À medida que o usuário vai realizando gestos a aplicação vai reconhecendo os dedos apresentados. Ao reconhecer certa quantidade de dedos é mostrado o número correspondente, como mostra a Fig. 6.1.

Figura 6.1. Contagem dos dedos

A aplicação reconhece no máximo cinco dedos em cada mão. A taxa de acerto chega a 95%, sendo a maioria dos erros devido a fusão de dois dedos adjacentes.

7. Conclusão e Trabalhos Futuros

Embora o algoritmo funcione realizando a contagem dos dedos apresentados, a aplicação pode funcionar de outras formas, como a apresentação aleatória de números na tela, aguardando que o usuário reproduza o mesmo com os dedos das mãos. O mesmo pode ser feito com o uso de voz, apresentando um comando com o som do número, ou mesmo realizando uma operação de soma simples. O uso contínuo do algoritmo pelo usuário faz com que ele naturalmente, observe os casos em que há erro de detecção e realiza o correto posicionamento e abertura entre os dedos das mãos, diminuindo a oclusão. O erro na detecção pode ser reduzido com o tratamento de vários

frames seguidos com uma máquina de estado que consiga assegurar o estado estável mais provável.

Trabalhos futuros devem realizar a identificação da posição das mãos na cena, tirando a restrição de o usuário estar em determinada distância do sensor. Sem essa restrição, pode-se identificar mais de um usuário no cenário o que possibilitaria o uso interativo para aplicação em torneios cooperativos.

8. Referências

R, Zhang and Y, Junsong. (2011) Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth Camera. In: Proceedings of the 19th ACM international conference on Multimedia, pages 1093-1096.

J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan. (2011) Vision-based hand-gesture applications. In Communications of the ACM, v.54:60–71.

A. Rodrigo, A. Jefferson & M. Francisco. (2012) AlfabetoKinect: Um aplicativo para auxiliar na alfabetização de crianças com o uso do Kinect. In SBIE, Anais.

Perrenoud, P. (2011) A pedagogia na escola das diferenças: fragmentos de uma sociologia do fracasso. In Porto Alegre: Artmed.

Amaro, D. G. and Macedo, L. (2011) Da lógica da exclusão à lógica da inclusão: reflexão sobre uma estratégia. In Seminário Internacional Sociedade Inclusiva, Anais. Belo Horizonte.

Jensen, M. B. (2011) Natural user interfaces from all angles: An investigation of interaction methods using depth sensing cameras. Aalborg University.

C. Chua, H. Guan, and Y. Ho. (2002) Model-based 3d hand posture estimation from a single 2d image. Image and Vision Computing, 20:191 – 202.

B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. (2003) Filtering using a tree-based estimator. In Proc. of IEEE ICCV.

S. Jaiswal, S. Bhadauria, R. S. Jadon, and T. Divakar. (2011) Brief description of image based 3d face recognition methods. 3D Research , 1(4):1-14.