



**Universidade Federal de Uberlândia
Faculdade de Engenharia Elétrica
Graduação em Engenharia Biomédica**

EDUARDO BORGES GOUVEIA

**IDENTIFICAÇÃO DE OBJETOS UTILIZANDO VISÃO
NEUROMÓRFICA E REDES NEURAIS CONVOLUTIVAS**

Uberlândia
2019

EDUARDO BORGES GOUVEIA

**IDENTIFICAÇÃO DE OBJETOS UTILIZANDO VISÃO
NEUROMÓRFICA E REDES NEURAIS CONVOLUTIVAS**

Trabalho apresentado como requisito parcial de avaliação
na disciplina Trabalho de Conclusão de Curso de
Engenharia Biomédica da Universidade Federal de
Uberlândia.

Orientador: Alcimar Barbosa Soares

Assinatura do Orientador

Uberlândia
2019

Dedico este trabalho à minha mãe por toda sua resiliência mesmo no mais sombrio dos tempos.

AGRADECIMENTOS

Ao professor Alcimar e ao Andrei por todo conhecimento compartilhado, compreensão e amizade nessa longa caminhada.

Aos meus companheiros de laboratório Eber, Ronaldo e Vinicius que sempre me inspiraram a ser uma pessoa melhor.

A minha irmã Jéssica que sempre cuidou de mim e me fez companhia.

Ao Laboratório de Engenharia Biomédica da Universidade Federal de Uberlândia (BioLab) pela infraestrutura e materiais que foram indispensáveis para a produção desse trabalho.

RESUMO

O reconhecimento e identificação de objetos através de visão computacional, usando algoritmos de aprendizagem profunda, são capazes de proporcionar resultados acurados, porém é um processo desafiador, e de alto custo computacional. Esse processo custoso é consequência de um conjunto de características das imagens tradicionais como, por exemplo, seu alto volume de dados, a resolução espacial das imagens que é tão maior quanto sua quantidade de pixels e a arquitetura robustas de hardwares de alto desempenho. Entretanto, o sistema visual biológico realiza essa tarefa com extrema maestria, sofisticação, alto desempenho e baixo custo energético; sua alta capacidade é fruto de anos de evolução e aperfeiçoamento. A engenharia neuromórfica utiliza uma abordagem totalmente bioinspirada para tentar alcançar a eficiência da natureza biológica e, conseqüentemente, abranger as fronteiras do conhecimento da neurociência. Aliando um sensor de visão neuromórfico com algoritmos bioinspirados, desenvolvemos um sistema de visão computacional para classificação e identificação de objetos utilizando redes neurais convolutivas. O modelo de aprendizado desenvolvido apresentou uma acurácia de 75.31% utilizando um *dataset* de imagens gravados por uma câmera neuromórfica (DVS-128).

ABSTRACT

The recognition and identification of objects through computer vision, using deep learning algorithms, are able to provide accurate results, but it is a challenging process, and a high computational cost. This costly process is a consequence of the characteristics of traditional images, such as high dimensionality, the spatial resolution of images that is as large as their amount of pixels and the robust architecture of hardware that require high performance. However, the biological visual system accomplishes this task with extreme mastery, sophistication, high performance and low energy cost; its high capacity is the fruit of years of evolution and improvement. Neuromorphic engineering uses a totally bioinspired approach to try to achieve the efficiency of biological nature and hence embrace the frontiers of knowledge in neuroscience. Combining a neuromorphic vision sensor with bioinspired algorithms, we developed a computer vision system for classification and identification of objects using convolutional neural networks. The developed learning model presented an accuracy of 75.31% using a dataset of images recorded by a neuromorphic camera (DVS-128).

LISTA DE ILUSTRAÇÕES

Figura 1: Disposição e detalhamento das camadas da retina humana, adaptado de Guyton (HALL, 2011).....	15
Figura 2: Absorção da luz pelo pigmento dos cones e bastonete, adaptado de Guyton (HALL, 2011).....	16
Figura 3: Curva de adaptação de sensibilidade à luz dos bastonetes e cones, adaptado de Guyton (HALL, 2011).....	16
Figura 4: Ilustração da transmissão de informação do córtex visual primário para as demais áreas visuais secundárias, adaptado de Guyton (HALL, 2011).....	17
Figura 5: Comparação entre imagens gravadas com sensores de visão tradicionais e sensores de visão neuromórficos, A esquerda podemos ver uma caneca gravada com um sensor de visão tradicional e a direita a mesma caneca gravada com um sensor de visão neuromórfico	19
Figura 6: Camada de Max Pooling com filtro 2x2 aplicado a imagem da esquerda. O processo ilustra o funcionamento de uma camada Max Pooling que, ao processar os dados, retira o maior valor do pixel dentro de seu filtro e atribui esse valor à um novo pixel que se.....	22
Figura 7: Imagem adaptada do trabalho de Gehrig (GEHRIG, 2018). Em (C) tem-se um disco que gira em sentido horário. Nesse disco é possível visualizar um pequeno ponto preto que, ao se deslocar com o movimento giratório do disco, produz variação de intensidade de.....	24
Figura 8: A imagem ilustra o processo de variação de luminosidade nas áreas adjacentes de um objeto em movimento na direção horizontal. Enquanto o objeto se move para a direita a luminosidade aumenta na região subsequente a direção do movimento, enquanto na área.....	25
Figura 9: Objetos utilizados para treino do modelo de aprendizagem profunda. São eles: caixa (A), caneca (B), celular (C), cubo (D), estilete (E), grampeador (F), lapiseira (G), mouse (H), tesoura (I). A esquerda de cada é possível visualizar a imagem gravada pelo sensor de visão neuromórfico e a direita a imagem do objeto gravado por um sensor de visão tradicional.	26
Figura 10: A imagem, na parte inferior, representa o esquema do posicionamento da câmera em relação aos objetos para gravação dos mesmos, a câmera foi posicionada a uma altura relativa ao objeto de 13 cm e distante, horizontalmente, 22.5 cm da base do objeto. A imagem superior é uma foto do resultado final do estúdio montado, onde é possível visualizar o sensor de visão neuromórfico DVS 128 e o objeto sendo gravado (grampeador), o servomotor foi acoplado ao fim da haste do tripé de apoio da câmera e não é visível na imagem superior, da mesma maneira a base giratória não é visível, pois não era de interesse que o sensor de visão neuromórfico captasse a variação de luminosidade da base. Acima da DVS-128 foi posicionado uma fonte de luz fluorescente, para diminuir a influência da variação de luminosidade, proveniente do ambiente externo, na gravação do objeto.	27
Figura 11: Arquitetura da rede neural convolutiva desenvolvido. Da esquerda para a direita observa-se a entrada da rede, a primeira camada convolutiva (que resulta em 32 atributos de tamanho 125x125), uma camada de pooling máximo (resultando em uma diminuição da dimensionalidade dos atributos para 61x61), uma camada convolutiva (que resulta em 86 atributos 60x60), seguido de um processo de pooling (que reduz o tamanho dos atributos para 29x29). Outra convolução, seguida de pooling (gerando 64 atributos de 28x28 e posteriormente	

13x13). Por fim, duas camadas totalmente conectadas, de tamanhos 64x1 e 9x1 consecutivamente.	29
Figura 12: Diversas representações da mesma gravação de um sensor de visão neuromórfico (DVS-128) sendo analisado dentro de Δt diferentes. Da esquerda para a direita temos períodos de integração de eventos de 10, 100, 200 e 500 ms respectivamente.	30
Figura 13: Imagem demonstrativa da diferença na representação de objetos gravados utilizando a DVS-128 que passaram por um processo de integração de eventos utilizando janelas fixas de eventos por frame. Da esquerda para a direita tem-se 1.000, 10.000 e 50.000 eventos/frame. De cima para baixo tem-se a gravação de um celular, um grampeador e um mouse.	32
Figura 14: Matriz de confusão resultante do processo de validação da rede neural convolutiva	34
Figura 15: Resultado da identificação e classificação do objeto pelo modelo de aprendizado profundo desenvolvido no trabalho. É possível perceber que o modelo se engana ao classificar o mouse (e) como um grampeador e, apesar de conseguir identificar o objeto na cena, por vezes, ele não consegue determinar todo o espaço ocupado pelo objeto (c)(a)(d).	36

LISTA DE TABELAS

Tabela 1: Gravações do sensor de visão neuromórfico DVS-128, separadas por classe , com seu respectivo tempo de duração que representa 360° de cada classe girando em torno de seu próprio eixo a uma velocidade de 2 rpm.....30

LISTA DE ABREVIATURAS E SIGLAS

ILSVRC – Imagenet Large Scale Visual Recognition Challenge

CNN – Rede Neural Convolutiva

SVM – Máquinas de Vetores de Suporte

DVS128 – Dynamic Vision Sensor

GPU – Unidades de Processamento Gráfico

SNN – Spiking Neural Networks

IBM – International Business Machines

SUMÁRIO

1. INTRODUÇÃO	12
2. REVISÃO BIBLIOGRÁFICA	15
2.1 A Visão BIOLÓGICA	15
2.2 A Visão NEUROMÓRFICA.....	18
2.3 Visão COMPUTACIONAL	20
3. METODOLOGIA DE DESENVOLVIMENTO	23
3.1. A REPRESENTAÇÃO NEUROMÓRFICA	23
3.2. GANHANDO FORMA	25
3.3. ATRIBUINDO SENTIDO	28
4. RESULTADOS E DISCUSSÃO	29
5. CONCLUSÕES	37
6. REFERÊNCIAS	39

1. Introdução

Os sistemas sensoriais possuem uma relevância vital no processo evolutivo. Eles exercem o papel de converter as entradas sensoriais em representações internas. Essa resposta do sistema sensorial deve ser rápida para se tornar uma eficiente ferramenta biológica de resposta a estímulos externos (SMITH, 2008). O grande desafio ao se desenvolver sistemas sensoriais artificiais é conseguir atingir o nível de eficiência, adaptabilidade e flexibilidade dos sistemas sensoriais biológicos.

O nome Neuromórfico é atribuído a sensores analógicos inspirados por sistemas biológicos responsáveis por processar informações sensoriais (LIU, 2010). A visão, por sua vez, assume um papel sensorial significativo no processo de interpretação de estímulos. Alguns desafios encontrados no processamento das imagens provenientes dos sensores de visão tradicionais são o alto custo computacional e a extração de informações relevantes através de algoritmos. Esse processo custoso é consequência de um conjunto de características das imagens tradicionais como, por exemplo, seu alto volume de dados, a resolução espacial das imagens que é tão maior quanto sua quantidade de pixels, o atributo de cor nas imagens (proveniente de três canais de informação) o que aumenta em três vezes a quantidade da informação a ser processada, por imagem. Todos esses fatores contribuem ao aumento da dimensionalidade da imagem que, por consequência, influencia no desempenho de algoritmos de processamento de imagens.

Entretanto, na natureza, podemos encontrar algumas soluções sofisticadas para processamento visual. O que ainda permanece como um desafio para a robótica e para inteligência artificial. Uma abordagem existente é a de investigação de sistemas biológicos para inspiração de sistemas artificiais.

Franceschini (FRANCESCHINI, 1992) propõe uma investigação acerca do sistema de controle visuomotor de insetos para o desenvolvimento de tecnologias bioinspiradas. O objetivo de Franceschini foi desenvolver um sistema visual bioinspirado em insetos voadores que, mesmo sem córtex visual, conseguem atingir altos níveis de performance em atividades como voar pelo ambiente, evitar obstáculos, se aproximar de objetos, empreender fuga de predadores ou pousar, executando tais tarefas com eficiência. Um exemplo que pode ser facilmente mencionado são as libélulas, abelhas e moscas que depositam sua confiança, primordialmente, no sistema visual para realizar essas tarefas e as realizam com extrema elegância.

Nossa atual inabilidade de desenvolver sistemas equiparados com a natureza se dá, em parte, pelo fato de que os sistemas robóticos atuais possuem grande dificuldade para interpretar as informações de maneira rápida, eficiente e fácil. A engenharia neuromórfica sugere uma

abordagem alternativa na busca pelo aprimoramento dos sistemas atuais. O desenvolvimento de sistemas neuromórficos encontra inspiração na capacidade do cérebro de resolver problemas complexos e, conseqüentemente, busca entender o funcionamento das complexas estruturas neuronais presentes nos seres vivos (LIU, 2015). Nesse contexto a engenharia neuromórfica busca construir sistemas tão eficientes e sofisticados quanto os encontrados na natureza (LIU, 2010), (BOAHEN, 2005), (DOUGLAS, 1995).

Rodney Douglas (DOUGLAS, 1995), um dos pioneiros no estudo de sistemas neuromórficos, exemplifica muito bem a significância desse campo de estudo por dizer que ela funciona como uma grande ponte entre os processos físico-químicos de um neurônio e sua significância computacional e permite com que nosso conhecimento sobre neurociência seja aplicado em dispositivos eletrônicos que responde a estímulos do mundo real da mesma maneira que o sistema nervoso biológico.

Os sensores neuromórficos visuais possuem circuitos que trabalham de maneira similar a retina. Na retina o processamento visual inicial é realizado pelos cones e bastonetes (fotorreceptores) e neurônios organizados de uma maneira que preserva a topografia (HALL, 2011). Circuitos neuromórficos tem uma organização física similar: fotorreceptores, elementos de memória e módulos computacionais são combinados em circuitos locais que processam, em tempo real, diferentes tipos de variações de luminosidade no tempo e espaço (LIU, 2015). A arquitetura computacional, tamanho compacto, e baixo consumo de energia de sensores neuromórficos fazem deles atrativos para a construção de sistemas sensoriais artificiais que tentam simular o processo biológico (LIU, 2015).

Para tratar do processamento da informação visual é preciso falar sobre visão computacional. Campo de estudo que tem ganhado muita força nos últimos anos. A *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* (DENG, 2009) é uma competição que acontece desde 2010 e é onde os competidores aplicam seus algoritmos de visão computacional na base de dados de imagens da ImageNet (que conta com mais de catorze milhões de imagens divididas em mais de vinte mil categorias) e competem para atingir a mais alta acurácia na tarefa de reconhecimento de objetos. Em 2011 o melhor resultado apresentava uma taxa de erro de 25% até que em 2012 uma rede neural convolutiva (CNN) diminuiu a taxa de erro para 16% (KRIZHEVSKY, 2012).

Nos anos seguintes muitos competidores continuaram tentando melhorar os métodos de classificação utilizando métodos convencionais de aprendizado de máquina, como a utilização de máquinas de vetores de suporte (SVM), porém a grande diminuição na taxa de erro em 2012 direcionou a atenção de diversos pesquisadores, e empresas, para o desenvolvimento de redes

neurais artificiais, mais precisamente modelos de aprendizagem profundo. Em 2017, impulsionada pela utilização de unidades de processamento gráfico (GPU), as redes neurais convolutivas estavam sendo amplamente utilizadas entre os competidores e a taxa de erro havia caído ainda mais, encontrando-se abaixo de 5%.

Tendo em vista a eficiência dos dispositivos neuromórficos e o alto desempenho da inteligência artificial em classificação de imagens, nosso objetivo principal é aliarmos a potencialidade do sensor de visão neuromórfica com a alta eficiência das redes neurais profundas, através da utilização de redes neurais convolutivas para criar um sistema bioinspirado de classificação e identificação objetos. Além de iniciar a construção de um *dataset* de gravações neuromórficas, para contribuir com a comunidade de pesquisa em visão neuromórfica.

Para tanto foi utilizado o sensor de visão neuromórfico Dynamic Vision Sensor (DVS128), criado pela *iniVation*, um instituto de informática de Zurique dedicado a criação de sistemas neuromórficos visuais. A DVS128 possui resposta baseada em eventos esparsados no tempo e no espaço, onde cada evento carrega 3 informações, a informação espacial de posicionamento da ocorrência do evento, a informação de diminuição ou aumento da luminosidade no ponto e o instante temporal que ocorreu o evento. Porém, além de possuir resposta baseada em eventos, a DVS128 conta com um processo assíncrono de envio da informação registrada, processo que diminui a redundância das informações e aumenta a resolução temporal da gravação (JIMENEZ-FERNANDEZ, 2010), (BARUA, 2016).

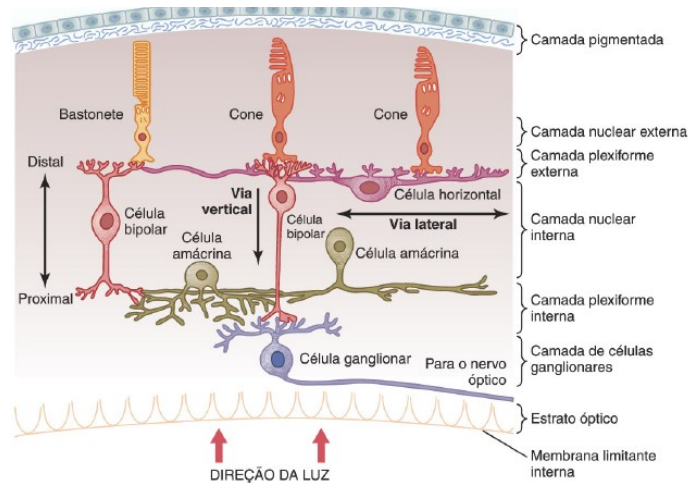
Por fim, foi utilizada a gravação da DVS128 para treinar o modelo de aprendizado desenvolvido nesse trabalho, dessa maneira conseguimos alinhar tanto hardware quanto software bioinspirados para construir um sistema de identificação de objetos.

2. Revisão bibliográfica

2.1 A Visão Biológica

A retina humana é um sistema complexo composto por três camadas principais, a primeira é a camada nuclear externa, a segunda é conhecida como nuclear interna e a terceira é a camada plexiforme interna, a Figura 1 mostra o esquema de distribuição de todas as camadas da retina. Cada qual com dezenas de milhares de células responsáveis por conduzir, e pre-processar os sinais visuais (KUFFLER, 1953), (MASLAND, 2001).

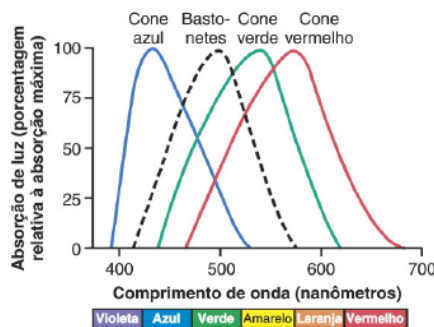
Figura 1: Disposição e detalhamento das camadas da retina humana, adaptado de Guyton (HALL, 2011)



A camada nuclear externa é constituída pelos cones e bastonetes. Os cones são as células capazes de reconhecer as cores, enquanto os bastonetes são células que estão fortemente ligadas a detecção de penumbra e principalmente são responsáveis pela visão em preto e branco. Existem três tipos de cones, cada um responsável por discriminar uma das três cores primárias.

Existem os cones mais sensíveis ao comprimento de onda de 400 - 530 nanômetros (sensíveis ao pigmento azul). Os cones que são mais sensíveis ao comprimento de onda de 440 a 610 nanômetros (responsáveis pela identificação da pigmentação verde) e os cones que são estimulados pelo comprimento de onda de 480 a 680 nanômetros (que por sua vez identificam os pigmentos vermelhos). Os três cones trabalhando juntos promovem a distinção de quase todas as graduações de cores, processo conhecido como mecanismo tri cromático de detecção de cores (RODIECK, 1988). A variação do comprimento de onda de acordo com a estimulação dos cones pode ser vista na Figura 2.

Figura 2: Absorção da luz pelo pigmento dos cones e bastonete, adaptado de Guyton (HALL, 2011)



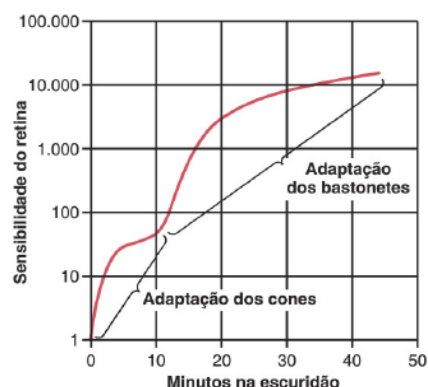
Já os bastonetes, por sua vez, são responsáveis pela detecção dos níveis de luminosidade. Eles são cerca de 100 vezes mais sensíveis que os cones, porém sua capacidade de detecção de luminosidade se restringe apenas a tons de cinza. Esses fotorreceptores também são responsáveis pelo fenômeno de adaptação ao escuro (WERBLIN, 1969). Os bastonetes sofrem um processo de adaptação de sensibilidade de acordo com a luminosidade do ambiente. Processo ilustrado na Figura 3.

Na camada nuclear interna é onde reside as células bipolares que são as responsáveis por intermediar a comunicação entre a camada nuclear externa e a camada plexiforme interna, transmitindo os sinais dos bastonetes e cones até as células ganglionares e amácrinas.

Porém, muito mais do que simplesmente intermediar a comunicação entre a camada nuclear externa e a camada plexiforme interna, as células bipolares, em colaboração com as células horizontais, exercem uma segunda atividade com grande importância evolutiva. Ao receber as informações dos cones e bastonetes e envia-las as células ganglionares, esse conjunto de células executam um somatório de luminosidade espacial elevando o limiar de despolarização das células da retina para um nível basal médio proporcional a quantidade de luz (PERKINS, 2018). É através desse fenômeno que somos capazes de enxergar em diferentes condições de luminosidade, necessitando apenas de alguns instantes de adaptação a luminosidade, além de diminuir a redundância das informações a serem processadas pelo córtex visual.

Um processo similar ao de somatório de luminosidade ocorre no processo de percepção de contraste. O somatório temporal é proveniente de alterações de luminosidade em pontos específicos da retina ocorrendo sucessivamente. Essa constância de mudança de luminosidade gera um processo excitatório ou inibitório. Se excitatório, existe uma percepção de aumento de luminosidade em um ponto específico, enquanto o processo inibitório gera uma sensação de

Figura 3: Curva de adaptação de sensibilidade à luz dos bastonetes e cones, adaptado de Guyton (HALL, 2011)



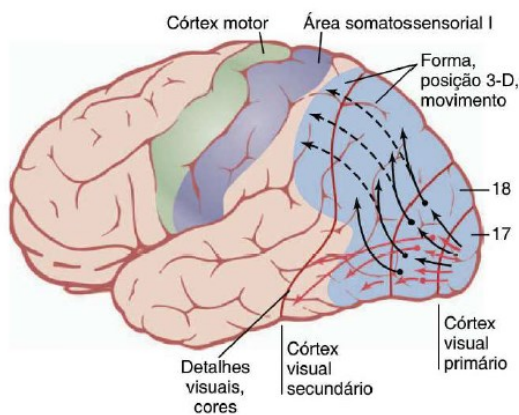
diminuição de luminosidade (SHARPE, 1993), (BARLOW, 1953), (SCHADE, 1956), (ENROTH-CUGELL, 1966).

Logo, quando olhamos para uma imagem com fundo preto e apenas um objeto branco no seu centro teremos um somatório da luminosidade da cena enxergada, e, se o objeto se mover na cena, teremos um aumento da luminosidade do ponto subsequente na direção do movimento enquanto haverá uma diminuição da luminosidade no ponto em que ele se encontrava no momento anterior.

Uma vez que essas informações chegam às células amácrinas e ganglionares elas são retransmitidas ao nervo óptico, sendo essas células as responsáveis por enviar os sinais da retina para o nervo óptico, que por sua vez carrega as informações até o córtex visual.

O córtex visual se divide em duas grandes áreas corticais. O córtex visual primário e as áreas de associação visual, que também são conhecidas como áreas visuais secundárias. São nas áreas de associação visual que ocorrem a análise dos significados visuais. (HALL, 2011).

Figura 4: Ilustração da transmissão de informação do córtex visual primário para as demais áreas visuais secundárias, adaptado de Guyton (HALL, 2011)



Na Figura 4 é possível ver a indicação, através das setas pretas contínuas, do fluxo da informação visual para uma região das áreas visuais secundárias. Essa região é responsável pela análise da posição de objetos visuais em relação ao corpo, pela forma física grosseira da cena visual e também do movimento do objeto nas cenas. Ou seja, essa região é capaz de nos apresentar informações de posição dos objetos, se eles estão em movimento ou não, e ainda classificar formatos (HALL, 2011).

2.2 A Visão Neuromórfica

A tecnologia inspirada na natureza tem cativado o interesse de diversos pesquisadores e centros de pesquisa. A chamada engenharia neuromórfica tem como objetivo desenvolver tecnologias bio e neuro-inspiradas. A missão é alcançar a capacidade de processamento biológico com seu baixo consumo de energia.

Sua natureza bioinspirada é a promessa para o desenvolvimento tecnológico da próxima geração de Inteligência Artificial. Empresas como a *International Business Machines* (IBM) e a Intel vem depositando sua confiança na eficiência energética e na capacidade dos dispositivos neuromórficos de abstrair informações de maneira não redundante.

Com foco em estabelecer uma comunidade de pesquisa em tecnologia neuromórfica a *Intel Labs* lançou a *Intel Neuromorphic Research Community* (INTEL, 2017) um esforço coletivo para estimular o governo, as indústrias, e os centros de pesquisa a desenvolver tecnologias neuromórficas e, assim, ampliar os horizontes da tecnologia bioinspirada.

A *IBM Research in Zurich* também vem trabalhando constantemente para ampliar a barreira do desenvolvimento de tecnologia artificial utilizando dispositivos neuromórficos. Sua missão é desenvolver novos materiais, novos equipamentos eletrônicos e novos sistemas computacionais neuromórficos.

Com todo seu esforço a IBM já alcançou grandes feitos. Sua pesquisa com dispositivos neuromórficos se tornou estado da arte em reconhecimento de gestos (ANDREOPOULOS, 2018). Um trabalho desenvolvido pelo pesquisador Andreopoulos e sua equipe na IBM, em 2018.

Para alcançar tal feito Andreopoulos e seu time utilizou dois sensores de visão neuromórfica (DVS), para criar um sistema de visão estéreo, e as chamadas *Spiking Neural Networks* (SNN) que são, como o nome sugere, redes neurais baseadas em *spikes*.

Inspirados pela maneira como os neurônios funcionam, os engenheiros neuromórficos, optaram por alterar radicalmente a abordagem da aquisição de dados sensoriais. Os sensores de visão tradicionais se baseiam em frames para realizar a aquisição de dados, enquanto os dispositivos neuromórficos utilizam a abordagem de eventos assíncronos para cada pixel.

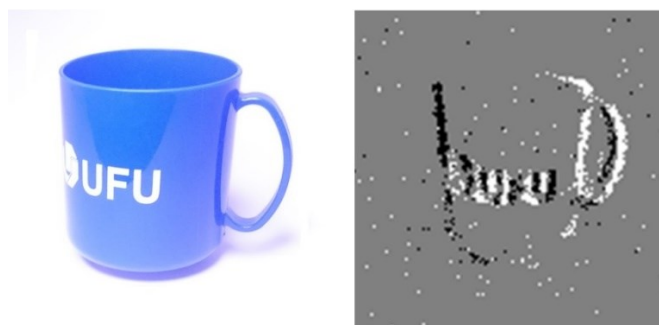
Nos dispositivos convencionais os valores dos pixels são capturados a cada frame, mesmo que a intensidade de luz no pixel permaneça inalterada. Esse processo gera redundância nos dados entregues por sensores de visão tradicionais, gerando uma grande quantidade de dados que para algumas aplicações são desnecessários. Um exemplo que pode ser citado é o de uma câmera de vigilância, onde os dados relevantes nessa aplicação seriam apenas aqueles que

carregam informações de movimento de objetos, pessoas, animais etc. Se tornando irrelevante as informações estáticas da cena.

Contudo os dispositivos neuromórficos, não gravam informações redundantes dado a sua natureza assíncrona, baseada em eventos. As retinas neuromórficas (como são conhecidas) enviam pacotes de dados informando apenas o endereço do pixel que captou uma variação significativa de luminosidade (POSCH, 2008), (LIU, 2015). Ou seja, quando algo na cena começa a se mover temos uma variação de luminosidade ao redor do objeto em movimento, processo muito similar ao que ocorre durante a estimulação das células da retina biológica. Esse processo gera uma diminuição na quantidade de dados necessários para carregar a informação do que está ocorrendo em uma cena, além de otimizar a transmissão de dados, consequência da sua natureza assíncrona (BRANDLI, 2014). Ou seja, existe uma diminuição significativa de redundância, enquanto preserva informações temporais de movimento.

Entretanto para ser possível obter uma representação visual das informações gravadas por um sensor de visão neuromórfica como a DVS (sensor de visão neuromórfica utilizada nesse trabalho) deve-se modelar os dados entregues por esses dispositivos para recriar a noção de frame. Para tanto é necessário agrupar os eventos recebidos durante certo período de tempo (Δt) e criar um pseudo frame. Na Figura 5 é possível observar a diferença entre a saída de dados de uma câmera tradicional baseada em frames (esquerda) e de uma Dynamic Vision Sensor (DVS), ambas gravando uma caneca.

Figura 5: Comparação entre imagens gravadas com sensores de visão tradicionais e sensores de visão neuromórficos, A esquerda podemos ver uma caneca gravada com um sensor de visão tradicional e a direita a mesma caneca gravada com um sensor de visão neuromórfico



Entretanto existem algumas considerações a serem feitas. Um sensor de visão neuromórfico estático, gravando uma cena estática, só apresentará registros ruidosos. Para que haja registro de informação visual relevante é necessário haver alteração de contraste na imagem sendo registrada. Alterações de contraste podem ser conseguidas com mudanças em

luminosidade ou ainda como consequência de movimento relativo entre o sensor neuromórfico e os elementos da cena.

Além dos já exemplificados trabalhos da IBM e Intel, existem outros centros de pesquisa que também utilizam o alto desempenho temporal da DVS, como o trabalho de Delbruck and Lichtsteiner em 2013 que utilizou a potencialidade da DVS para criar um goleiro robótico com tempo de resposta de 3 ms (DELBRUCK, 2013), o trabalho de Drazen em 2011 que utilizava a DVS para rastrear o movimento de partículas em um fluido (DRAZEN, 2011). O trabalho de Censi em 2013 que utilizou a DVS para aumentar a agilidade e o tempo de resposta de robôs voadores autônomos (CENSI, 2013). O trabalho de Chen em 2019, utiliza visão neuromórfica para identificar pedestres (CHEN, 2019).

2.3 Visão Computacional

Independente de se utilizar um sensor de visão convencional, ou um sensor de visão neuromórfico, ainda necessitaremos atribuir significado as informações gravadas através desses dispositivos. Tarefa que é incumbida a um campo de pesquisa chamado Visão Computacional.

Nos últimos dez anos as fronteiras da visão computacional têm estado em constante crescimento. Algoritmos mais sofisticados foram desenvolvidos que permitem classificar, detectar e ainda significar informações de uma cena (KRIZHEVSKY, 2012), (LIANG, 2019), (RINALDO, 2019), (SIMON, 2018), (SHAFIEE, 2017), (GRIFFIN, 2007). Entretanto, mesmo com tantos avanços e aprimoramentos na interpretação computacional das informações, esse segmento ainda se encontra, tecnologicamente, longe de alcançar a mesma eficiência, na interpretação das informações de uma cena, que uma criança de três anos de idade.

As atuais máquinas necessitam de bastante poder de processamento para realizar a tarefa de atribuir significado aos valores de pixels que representam uma imagem, enquanto crianças conseguem realizar essa tarefa facilmente, resultado do incrível trabalho da natureza em criar um sistema visual extremamente sofisticado. Para entender um pouco melhor como o corpo humano é capaz de realizar essa tarefa, que nossas mais sofisticadas máquinas ainda têm dificuldade em realizar, é preciso entender que o sistema visual contempla não somente um par de retinas, mas possui um sistema de processamento visual no cérebro que é tão sofisticado quanto os olhos.

Os métodos de aprendizado de máquina são diversos e consistem em ensinar ao computador a distinguir objetos. Para tanto é possível induzir o aprendizado através do que é conhecido como aprendizado supervisionado, que consiste em dar uma grande quantidade de amostras de características rotuladas com alguma informação útil acerca daquelas características.

O algoritmo deve ser capaz de aglomerar as características que compõe alguma informação e criar um classificador para essas características.

A partir desses classificadores o algoritmo deve ser capaz estimar o que um conjunto de características, nunca vistas pelo algoritmo antes, representa. Ou seja, se for ensinado a uma máquina que os atributos de uma cadeira são ter encosto, assento e quatro apoios, é esperado que o algoritmo seja capaz de determinar que um objeto que possui essas características é uma cadeira.

Entretanto, ao longo dos anos, a meta de se aprimorar a visão computacional continuou a inspirar muitos pesquisadores. O computador deve ser capaz de distinguir não somente uma cadeira de uma mesa, mas é necessário que o computador consiga diferenciar diversos objetos em cena e com um grau de sofisticação suficiente para distinguir lobos de cães, por exemplo.

Ao decorrer dos anos, de desenvolvimento da visão computacional, a *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)* (DENG, 2009) se tornou um grande motivador para equipes de pesquisadores, incentivando-os a alcançar cada vez melhores resultados na classificação de imagens. Durante muitos anos, foram utilizadas técnicas convencionais de aprendizado de máquina, como máquinas de vetores de suporte (SVM), Correlação Canônica, K-Nearest Neighbors (KNN) (CORNFIELD, 1972), (DEVIJVER, 1982), (FUKUNAGA, 2013). Porém, em 2012, um grupo de pesquisadores, contra todas as expectativas, utilizou uma rede neural convolutiva (CNN) para criar um classificador de imagens ganhando a competição, daquele ano, e diminuindo a taxa de erro para 16% (KRIZHEVSKY, 2012).

2.4 Redes Neurais Convolutivas

As redes convolucionais fazem parte do que é conhecido como de aprendizado profundo, são mecanismos de aprendizado de máquina que utilizam múltiplas camadas, cada qual responsável por aprender diferentes atributos, processo que é biologicamente inspirado e conhecido como aprendizado em cascata (WENG, 1992). O aprendizado em cascata tem sua inspiração nos modelos biológicos propostos por David H. Hubel e Torsten Wiesel das células visuais.

Ambos ganharam o prêmio Nobel de medicina e psicologia em 1981, por suas descobertas acerca do processamento da informação visual pelo cérebro. Em seus estudos eles demonstraram que a imagem é interpretada com base no seu contraste, formato e pelo movimento da imagem através da retina. Cada trecho da informação visual é passado de uma célula à outra na retina e cada tipo de célula é responsável por certos detalhes na imagem (HUBEL, 1979), (HUBEL, 1988).

Essa granularidade da interpretação da imagem é a base de inspiração para as redes convolutivas, onde cada camada da rede é responsável por determinados atributos, que por sua vez, quando combinados, representam uma informação.

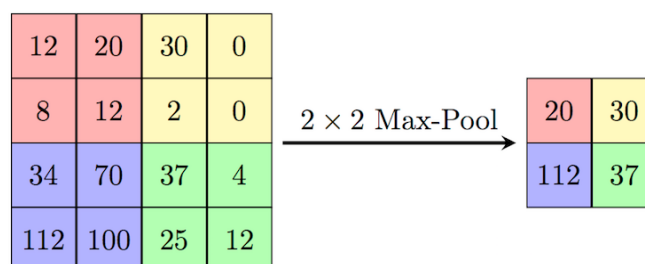
O uso das CNNs, principalmente aliado a tarefas de reconhecimento de imagens, se mostrou bastante eficaz e fez com que diversos pesquisadores direcionassem seus esforços para o desenvolvimento de redes convolutivas no processo de visão computacional. Alguns anos após as CNNs serem apresentadas na competição da *ImageNet*, todos os demais participantes já estavam apresentando diferentes modelos de CNNs com diferentes topologias que aperfeiçoaram o seu desempenho.

Comumente as redes neurais convolutivas são compostas por cinco diferentes tipos de camadas. Cada qual responsável por algum processo, como extrair atributos utilizando filtros iniciados com valores aleatórios (camadas convolutivas), diminuir a dimensionalidade dos atributos (camadas *Max Pooling*), evidenciar variações de contraste na imagem (camadas ReLu) e classificar os atributos como pertencentes a alguma classe (camadas totalmente conectadas e Softmax).

Cada filtro nas camadas convolucionais gera um atributo para a imagem. Em uma camada com n filtros, tem-se n atributos diferentes resultantes do processo de convolução do filtro com a imagem de entrada. Já as camadas ReLu tem a função de aumentar a não linearidade das imagens. Nessa camada, os valores zero, e negativos, das imagens são retirados, restando apenas valores positivos. Diminuindo, assim, as mudanças graduais das imagens e evidenciando as mudanças abruptas de cor (KUO, 2016), (HE, 2015).

As camadas Max Pooling são responsáveis por diminuir a dimensionalidade dos dados de entrada, através de um filtro com dimensionalidade configurável (MURRAY, 2014). Esse processo é realizado ao se retirar o maior valor de um grupo de pixels de tamanho igual ao do filtro de pooling desejado e substituir os valores desses pixels por apenas um pixel possuindo o valor máximo do grupo de pixels abrangidos pelo tamanho do filtro da camada de Max Pooling (Figura 6).

Figura 6: Camada de Max Pooling com filtro 2x2 aplicado a imagem da esquerda. O processo ilustra o funcionamento de uma camada Max Pooling que, ao processar os dados, retira o maior valor do pixel dentro de seu filtro e atribui esse valor à um novo pixel que se



Por fim tem-se uma camada totalmente conectada e uma camada Softmax. Ambas trabalham juntas no processo de classificação do resultado retornado pelas camadas convolucionais e camadas de pooling. Essas camadas são responsáveis por calcular a probabilidade de cada entrada pertencer a uma das classes. A camada totalmente conectada recebe esse nome pois cada pixel da camada anterior está ligado a todos os pixels da camada subsequente. Essa camada é responsável por classificar a imagem de entrada em cada uma das classes possíveis, enquanto a camada Softmax desempenha o papel de normalizar o resultado da camada totalmente conectada, criando, assim, a distribuição das probabilidades da imagem de entrada pertencer a cada classe.

3. Metodologia de Desenvolvimento

Com a união entre a bioinspiração dos sensores de visão neuromórficos e as Redes Neurais Convolutivas foi utilizado uma *Dynamic Vision Sensor* (DVS128) para gravar nove classes de objetos e criamos um modelo de rede neural para identificar e classificar esses objetos em cena.

3.1. A Representação Neuromórfica

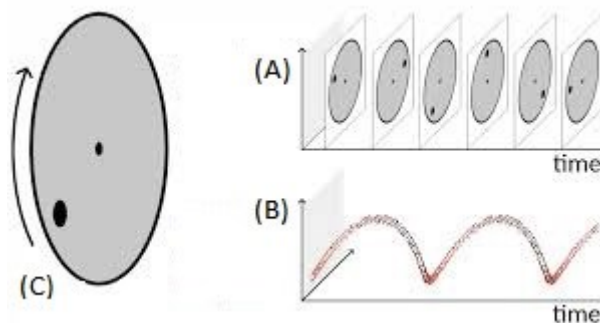
Diferentemente dos sensores de visão tradicionais que capturam frames, os sensores de visão neuromórficos utilizam uma abordagem distinta para gravar cenas. Em sensores de visão convencionais, a cena é gravada através do registro dos diversos níveis de intensidade de luz em cada pixel registrado, no instante de registro do frame. Através de uma junção de frames consecutivos, tem-se a noção temporal de eventos. Porém, esse processo é tão rápido quanto a capacidade da câmera de registrar um frame. Em um sensor de visão tradicional, comumente comercializado no mercado, teremos uma capacidade de cerca de 30 frames por segundo (FPS).

A concepção neuromórfica utiliza uma abordagem bioinspirada na aquisição de informações visuais, de aquisição de eventos de maneira assíncrona (POSCH, 2008). Semelhantemente as retinas biológicas, os sensores de visão neuromórficos, consideram a variação de intensidade luminosa como um evento contendo informação visual relevante. Nesse momento, é transmitido, de maneira assíncrona, o evento carregando as informações espaciais, temporais e de intensidade dessa variação de contraste.

Por se tratar de um método assíncrono de transmissão de informação, não há a formação de frames, ou seja, cada pixel dispara uma informação no momento em que é registrado uma

variação de luminosidade. Logo, ao utilizar sensores de visão neuromórficos, não existe a concepção de frames, em seu lugar é utilizado a concepção de eventos esparsados temporal e espacialmente.

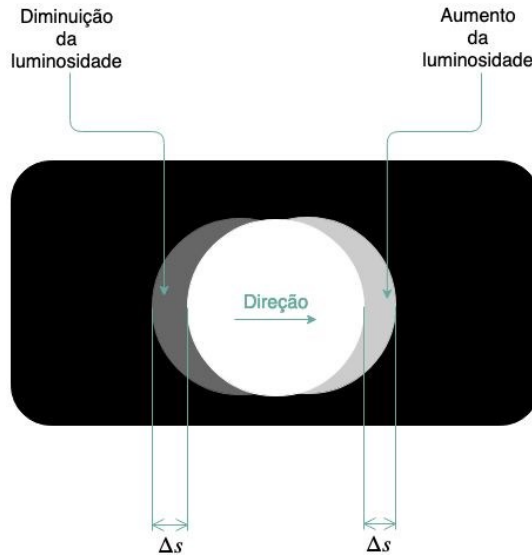
Figura 7: Imagem adaptada do trabalho de Gehrig (GEHRIG, 2018). Em (C) tem-se um disco que gira em sentido horário. Nesse disco é possível visualizar um pequeno ponto preto que, ao se deslocar com o movimento giratório do disco, produz variação de intensidade de



Com seu método assíncrono de aquisição, os sensores de visão neuromórficos possuem uma alta resolução temporal ($\sim 15 \mu s$), diferentemente dos sensores de visão tradicionais. Pode-se perceber, ilustrativamente, a diferença entre o método tradicional e o método neuromórfico de aquisição de informações visuais, ao analisar a Figura 7. Nela é possível perceber o rastreamento da posição do ponto, no disco giratório, mudando ao longo do tempo, e como os sensores de visão registram esse movimento.

Além do registro da variação de contraste no espaço tempo, a DVS-128, também registra se a variação da intensidade luminosa foi positiva ou negativa. Logo, se o objeto se desloca horizontalmente, a região adjacente ao movimento estará aumentando em intensidade luminosa, enquanto a outra região, horizontalmente oposta, estará diminuindo (Figura 8).

Figura 8: A imagem ilustra o processo de variação de luminosidade nas áreas adjacentes de um objeto em movimento na direção horizontal. Enquanto o objeto se move para a direita a luminosidade aumenta na região subsequente a direção do movimento, enquanto na área



Com as informações dos eventos sendo recebidas, pode-se recriar a noção de frame, agrupando os eventos em um determinado período de tempo, cada qual em sua posição espacial e distinguindo sua variação de luminosidade.

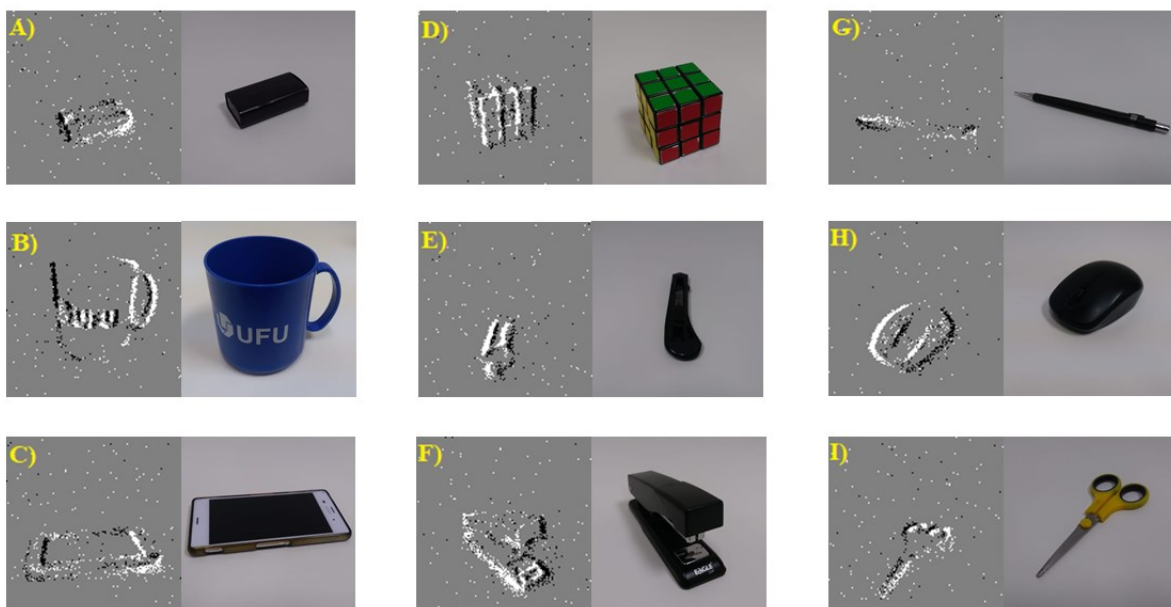
Para realizar esse processo foi decidido recriar os frames utilizando 8 bits. Aos eventos que se destinavam a diminuição de luminosidade foi atribuído o valor 0, enquanto os eventos que representavam um aumento de intensidade luminosa receberam o valor 255, onde não havia variação de luminosidade relativa atribuímos o valor de 128. Cada frame recriado, era resultado de uma integração dos eventos dentro de um período de 100ms, nesse processo o valor final do pixel era determinado pelo resultado da análise da quantidade de eventos positivos e eventos negativos dentro do período de tempo determinado.

3.2. Ganhando Forma

Para ensinar uma máquina que um gato é um gato, é necessário apresentar para a máquina diversas imagens de gatos, em diferentes posições, tamanhos e possuindo diferentes características. Esse processo de aprendizagem de máquina é conhecido como aprendizado indutivo. Nesse trabalho foram utilizados artefatos inanimados e comuns para serem objetos de aprendizado da máquina. Um modelo de rede neural convolutiva foi utilizado para aprender as características dos objetos gravados através da DVS 128.

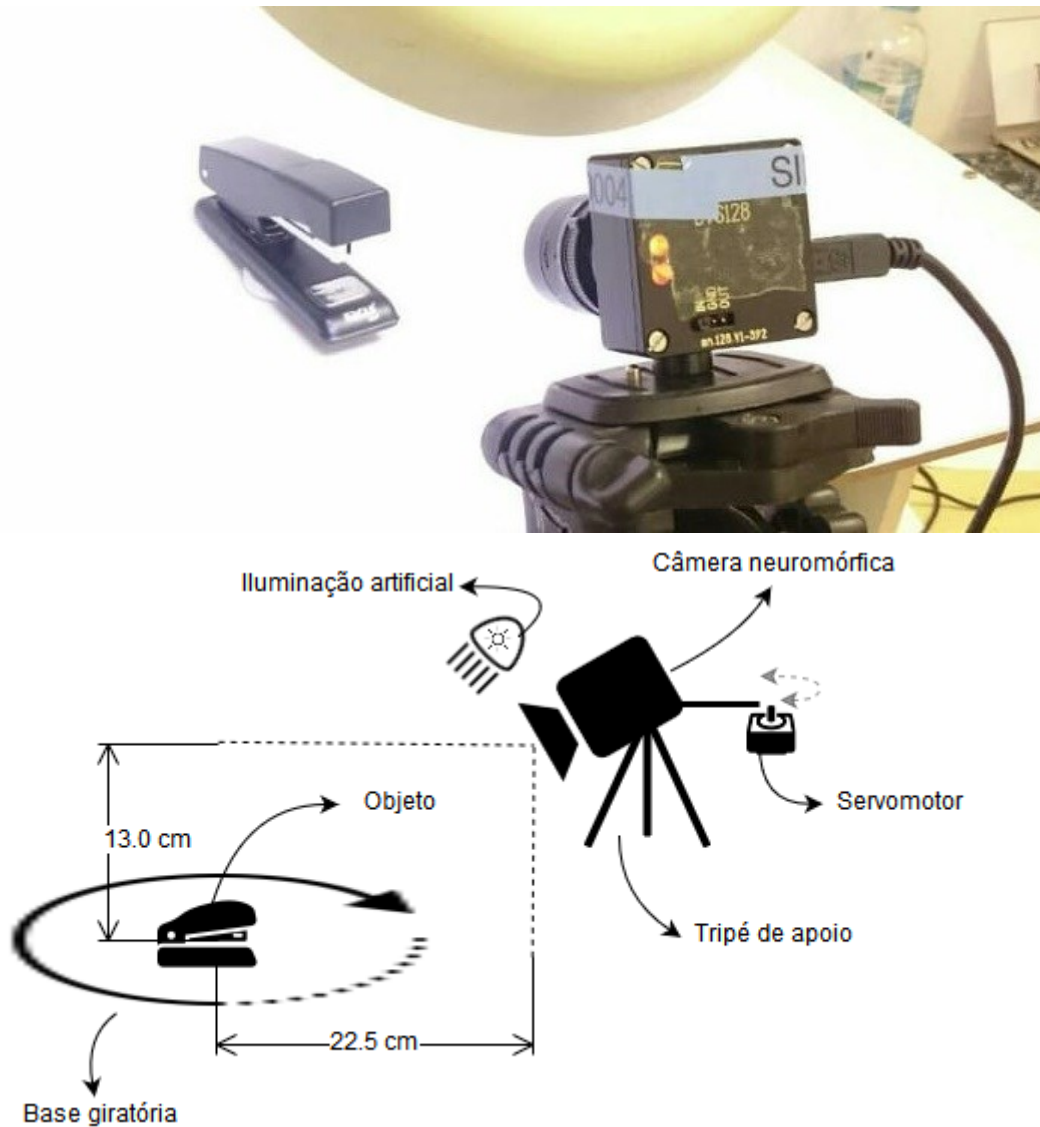
Como, em métodos indutivos, precisamos apresentar ao modelo diversos exemplos dos objetos, foi gravado cada objeto em diferentes ângulos de modo que foi possível obter um número de amostras suficientes para que cada objeto seja reconhecido mesmo que visto de lados diferentes. No modelo de aprendizado proposto nesse trabalho, foram utilizadas 9 classes como objetos de aprendizado (Figura 9).

Figura 9: Objetos utilizados para treino do modelo de aprendizagem profunda. São eles: caixa (A), caneca (B), celular (C), cubo (D), estilete (E), grampeador (F), lapiseira (G), mouse (H), tesoura (I). A esquerda de cada imagem é possível visualizar o objeto gravado pelo sensor de visão neuromórfico e a direita a imagem do objeto gravado por um sensor de visão tradicional.



Para conseguir as imagens de diversos ângulos, foi montado um estúdio com base giratória que girava a uma velocidade de 2 rpm, onde cada objeto foi posicionado sobre a base giratória, permitindo que o objeto fosse gravado em 360°. O objeto foi gravado durante 30 segundos, representando uma volta completa do objeto em torno do próprio eixo. A câmera foi posicionada em um apoio fixo à 22,5 cm de distância, na horizontal, da base do objeto e angulação de 30° em relação a base (Figura 10).

Figura 10: A imagem, na parte inferior, representa o esquema do posicionamento da câmera em relação aos objetos para gravação dos mesmos, a câmera foi posicionada a uma altura relativa ao objeto de 13 cm e distante, horizontalmente, 22.5 cm da base do objeto. A imagem superior é uma foto do resultado final do estúdio montado, onde é possível visualizar o sensor de visão neuromórfico DVS 128 e o objeto sendo gravado (grampeador), o servomotor foi acoplado ao fim da haste do tripé de apoio da câmera e não é visível na imagem superior, da mesma maneira a base giratória não é visível, pois não era de interesse que o sensor de visão neuromórfico captasse a variação de luminosidade da base. Acima da DVS-128 foi posicionado uma fonte de luz fluorescente, para diminuir a influência da variação de luminosidade, proveniente do ambiente externo, na gravação do objeto.



Entretanto, a base giratória sozinha não é suficiente para gravar os objetos com a DVS-128. Devido a sua natureza bioinspirada, é necessário que haja variação na intensidade de luz (como é discutido na sessão 2.2.1). Para tanto foi criado um mecanismo de translação horizontal da DVS-128. Nesse processo o sensor de visão neuromórfico realizou micro movimentos horizontais impulsionados por um servo motor acoplado à uma haste presa a base do sensor. Nesse processo a câmera neuromórfica se movia continuamente de um lado para o outro.

Esse processo de movimentação do sensor é semelhante ao movimento sacádico realizado pela retina humana. Processo em que os olhos se movimentam rápida, e simultaneamente, durante

o processo de fixação do olhar. Pois, assim como os sensores de visão neuromórficos, a retina humana não consegue visualizar cenas em estados fixos (CARPENTER, 1988).

Por fim, o *dataset* criado para esse trabalho, contabilizou 2700 imagens, sendo elas 300 imagens de cada uma das 9 classes escolhidas. Como o objeto foi gravado sobre uma base giratória e as imagens integradas em um período de tempo de 100 ms, as imagens constituintes de cada classe representam uma volta completa do objeto em torno do próprio eixo com um intervalo de 1,2 ° entre elas.

3.3. Atribuindo Sentido

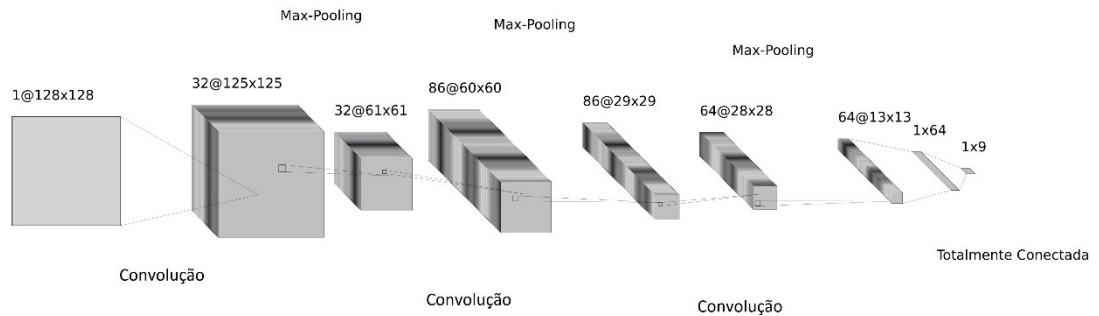
Assim como a retina, o sensor de visão neuromórfico não atribui significado ao que grava. É necessário o trabalho em conjunto entre a retina e o córtex visual para significar o que é visto. Nesse trabalho foi utilizado o sensor de visão neuromórfico em conjunto com uma rede neural, em um processo de aprendizado supervisionado, para realizar a tarefa de significar o que é gravado.

As técnicas de aprendizado profundo (ou *deep learning*) tem ganhado espaço em reconhecimento de imagens nos últimos anos, como já discutido anteriormente, entre elas podemos destacar as redes neurais convolutivas (CNN), que apresenta as melhores performances, atualmente, em reconhecimento de imagens (SIMON, 2018), (KRIZHEVSKY, 2012).

A arquitetura da CNN criada para esse trabalho contou com 3 camadas convolucionais, 3 camadas ReLu, 3 camadas Max Pooling, uma camada totalmente conectada e uma camada softmax. Na Figura 11 é possível visualizar o diagrama arquitetural do modelo criado.

Cada filtro nas camadas convolucionais gera um atributo para a imagem. Nesse trabalho, a primeira camada convolucional conta com 32 filtros 8x8, ou seja, para cada imagem de entrada tem-se 32 representações daquela imagem, só na primeira convolução. A segunda camada convolucional possui um total de 86 filtros 6x6, enquanto a última camada convolucional possui 64 filtros 6x6.

Figura 11: Arquitetura da rede neural convolutiva desenvolvido. Da esquerda para a direita observa-se a entrada da rede, a primeira camada convolutiva (que resulta em 32 atributos de tamanho 125x125), uma camada de pooling máximo (resultando em uma diminuição da dimensionalidade dos atributos para 61x61), uma camada convolutiva (que resulta em 86 atributos 60x60), seguido de um processo de pooling (que reduz o tamanho dos atributos para 29x29). Outra convolução, seguida de pooling (gerando 64 atributos de 28x28 e posteriormente 13x13). Por fim, duas camadas totalmente conectadas, de tamanhos 64x1 e 9x1 consecutivamente.



Como método de validação do modelo de aprendizado desenvolvido, foi utilizado o método *Holdout* no qual é separado uma porção do montante total de exemplos do *dataset* que é utilizada para treino e a porção restante é utilizada para teste (YADAV, 2016). Comumente são utilizados 2/3 do total de amostras no *dataset* para treino e 1/3 é utilizado para a validação do treino, por questões de aproximação foi utilizado 70% das amostras para treino e 30% para testes, o que representa 1890 imagens de treino e 810 imagens de teste, divididas igualmente entre as nove classes utilizadas.

4. Resultados e Discussão

Devido à natureza assíncrona de um sensor de visão neuromórfico, como a DVS-128, as gravações representam um aglomerado de informações no tempo. Cada uma das gravações realizadas durante a execução desse trabalho tem a duração de 30 segundos de eventos esparsos. Essas informações poderiam ser tratadas e manipuladas de diferentes maneiras, de acordo com o interesse do usuário. Porém, como a entrada esperada da CNN é uma imagem, foi então escolhido trabalhar com a integração dos eventos no tempo para formar um frame. Dessa maneira foi possível obter representações gráficas, semelhantes a imagens, das gravações realizadas com a DVS-128.

Esse processo resultou em um *dataset* de 300 imagens por classe, como foi utilizado 9 classes nesse trabalho o tamanho total do *dataset* foi de 2700 imagens (Tabela 1). Entretanto, o processo de conversão das informações de um sensor de visão neuromórfico poderia ter sido realizado de maneiras diversas (FRANCO, 2013). Outro método comum de realizar esse

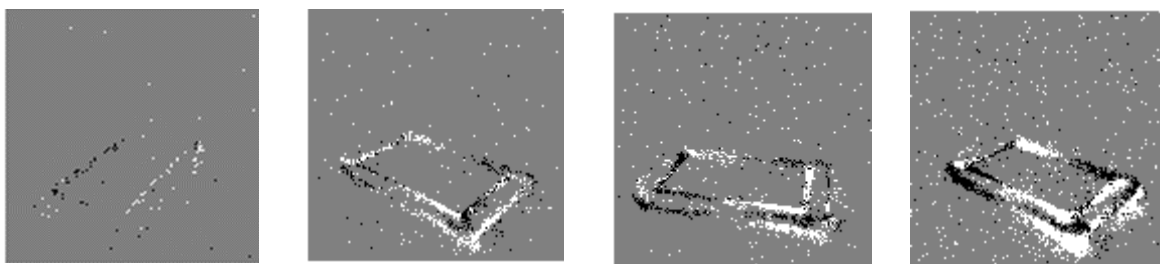
procedimento seria a integração dos eventos até que fosse atingido uma quantidade satisfatória para representar o objeto. Ou seja, antes de formar a imagem o algoritmo aguardaria o tempo que fosse necessário até que a quantidade de pontos desejada fosse atingida.

Tabela 1: Gravações do sensor de visão neuromórfico DVS-128, separadas por classe, com seu respectivo tempo de duração que representa 360° de cada classe girando em torno de seu próprio eixo a uma velocidade de 2 rpm.

Categoria	Duração [s]
Caixa	30
Mouse	30
Celular	30
Lapiseira	30
Estilete	30
Grampeador	30
Tesoura	30
Cubo	30
Caneca	30
Total	270

Esses métodos proporcionam percepções diferentes dos objetos em cena, e é possível que sem a quantidade suficiente de movimento relativo entre o objeto e a DVS-128 não haja disparos de eventos suficientes para exibir uma representação real do objeto gravado. Entretanto, como a representação, em frame, da gravação de um objeto utilizando um sensor de visão neuromórfico é modulada de acordo com o interesse do usuário, é possível que se integre os pontos durante tempo suficiente para que ele assuma uma representação mais fielmente próxima do objeto real. Foi mostrado, na Figura 12, a gravação de um objeto, onde o frame foi construído integrando os eventos gravados pelo sensor de visão neuromórfico, utilizando diversos intervalos de tempo.

Figura 12: Diversas representações da mesma gravação de um sensor de visão neuromórfico (DVS-128) sendo analisado dentro de Δt diferentes. Da esquerda para a direita temos períodos de integração de eventos de 10, 100, 200 e 500 ms respectivamente.

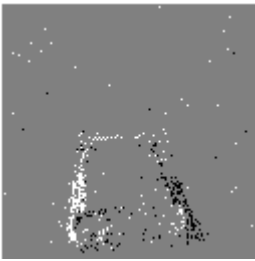
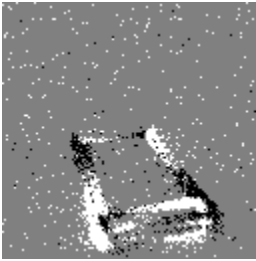
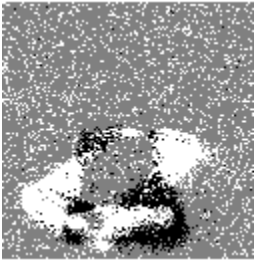
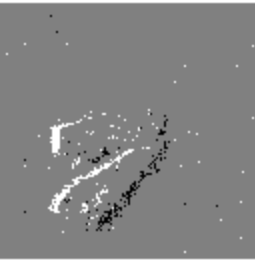
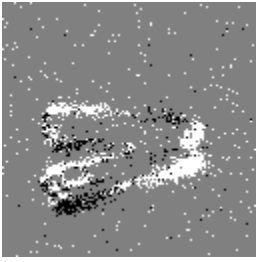
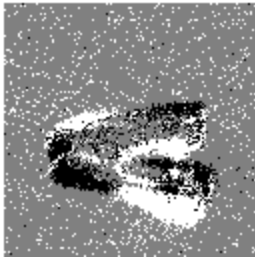
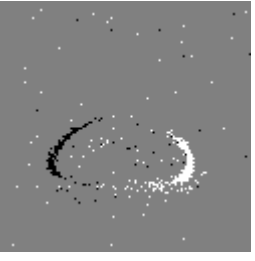
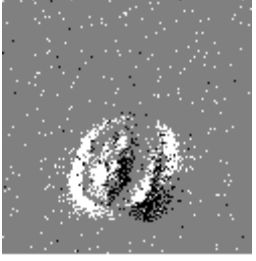
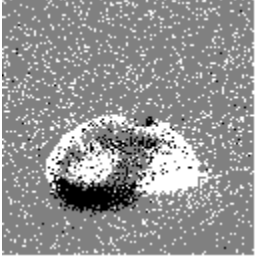


É possível perceber que, dependendo da quantidade de eventos integrados, a imagem percebida representa dois extremos, ou existe uma falta de informação suficiente para representar o objeto, ou existe tanta informação que dificulta a distinção do mesmo. Entretanto, a quantidade de eventos estimulados também é influência da quantidade de movimento relativo entre a DVS-128 e o objeto, o aumento de movimento relativo, causa maior variação de luminosidade, que consequentemente proporciona mais disparos de eventos.

Uma abordagem diferente poderia ser determinar um número fixo de eventos para cada frame, porém, não só a velocidade e o movimento que determinam a quantidade de eventos, o formato do objeto e a quantidade de arestas também influenciam no número de eventos.

A Figura 13 demonstra a diferença entre a representação de alguns objetos integrando os eventos em uma janela fixa de eventos. Na primeira coluna foi integrado 1.000 eventos por frame, enquanto na segunda e terceira foram integrados, 10.000 e 50.000 eventos por frame, respectivamente. É possível perceber que a mesma quantidade de eventos gera qualidades distintas nas representações de diferentes objetos.

Figura 13: Imagem demonstrativa da diferença na representação de objetos gravados utilizando a DVS-128 que passaram por um processo de integração de eventos utilizando janelas fixas de eventos por frame. Da esquerda para a direita tem-se 1.000, 10.000 e 50.000 eventos/frame. De cima para baixo tem-se a gravação de um celular, um grampeador e um mouse.

Objeto	Eventos/Frame		
	1.000	10.000	50.000
Celular			
Grampeador			
Mouse			

Logo, dado o ambiente controlado montado no laboratório, com velocidade de movimento da DVS-128 controlado e rotação do objeto a uma velocidade também determinada empiricamente, assumiu-se que a integração de eventos dentro do intervalo de 100 ms seria suficiente para uma representação de qualidade do objeto. Para diferenciar os objetos gravados foi necessário a criação de um modelo classificatório, nesse estudo foi utilizado uma rede neural convolutiva.

Inicialmente, foram realizados o treinamento e o teste da rede de aprendizado profundo, processo no qual se faz necessário provisionar, para o modelo de aprendizagem, tanto informações para treino, quanto informações para teste e validação do modelo criado.

Existem métodos diferentes de validação de um modelo de treinamento, um dos métodos, como exemplo, é conhecido como *k-fold* que consiste em separar o conjunto de dados em k

subconjuntos do mesmo tamanho e diferentes entre si. Dessa maneira um grupo é separado para ser o conjunto de imagens para teste/validação, enquanto os outros $k-1$ são utilizados para treinamento do modelo. Esse processo é realizado k vezes, sendo que o conjunto de teste varia entre cada repetição (KOHAVI, 1995).

Uma variação do k -fold é o método *leave-one-out*, nele tem-se que o k é igual à quantidade total de informações no *dataset* (N), onde são realizadas N cálculos de erros, um para cada imagem. Porém, essa abordagem requer um altíssimo custo computacional (SHAO, 1993).

O método de validação utilizado nesse trabalho foi o *holdout*, onde o *set* de treino consiste, comumente, em 2/3 de N , enquanto o *set* de teste/validação possui, comumente, 1/3 de N (YADAV, 2016). Esse método requer menos poder computacional, quando comparado com outros métodos; pois a validação é realizada somente uma vez, diferentemente do método k -fold onde o cálculo é realizado k vezes.

O *dataset*, criado nesse projeto, foi dividido em dois grupos, sendo 70% do total para o *set* de treinamento e 30% para o *set* de teste/validação. Com isso, foram utilizadas 1890 imagens para treino e 810 imagens para teste.

Uma matriz de confusão foi criada (Figura 14) para demonstrar a assertividade do modelo criado nesse trabalho. Nela é possível verificar que, em objetos como o cubo e o celular, o modelo de aprendizado exerce uma excepcional capacidade de dedução, alcançando 100% e 94.6% de acerto, respectivamente. Entretanto, objetos semelhantes em formato, como uma lapiseira e uma tesoura são mais facilmente confundidos entre si. Essa aparente tendência do modelo de confundir uma lapiseira e uma tesoura pode ser atribuída ao formato similar da topografia dos dois objetos, quando gravados pela câmera neuromórfica, além disso a baixa quantidade de exemplos de cada classe pode ter contribuído para a pouca capacidade do modelo de distinguir imagens com diferenças sutis entre a lapiseira e a tesoura. Também é possível que, em alguns ângulos, os objetos se assemelhem topograficamente. Já objetos como o cubo, são mais facilmente distinguíveis em suas diversas faces, o que torna sua classificação mais fácil. Por fim a acurácia geral, do modelo de aprendizado convolutivo, foi de 75.31%.

Figura 14: Matriz de confusão resultante do processo de validação da rede neural convolutiva

Acurácia: 75.31%

Caixa	86.7% 78	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	10.2% 9	3.8% 3
Caneca	0.0% 0	93.8% 90	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Celular	0.0% 0	0.0% 0	94.6% 88	0.0% 0	0.0% 0	0.0% 0	1.1% 1	1.1% 1	0.0% 0
Cubo	0.0% 0	0.0% 0	0.0% 0	100.0% 90	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Estilete	19.7% 23	0.0% 0	0.0% 0	0.0% 0	66.3% 61	0.0% 0	2.1% 2	1.1% 1	3.8% 3
Grampeador	0.0% 0	6.3% 6	0.0% 0	0.0% 0	0.0% 0	98.4% 60	3.2% 3	3.4% 3	23.1% 18
Lapiseira	0.9% 1	0.0% 0	1.1% 1	0.0% 0	13.0% 12	0.0% 0	53.7% 51	0.0% 0	32.1% 25
Mouse	12.8% 15	0.0% 0	0.0% 0	0.0% 0	3.3% 3	1.6% 1	0.0% 0	76.1% 67	5.1% 4
Tesoura	0.0% 0	0.0% 0	4.3% 4	0.0% 0	17.4% 16	0.0% 0	40.0% 38	8.0% 7	32.1% 25
	Caixa	Caneca	Celular	Cubo	Estilete	Grampeador	Lapiseira	Mouse	Tesoura

Classe Verdadeira

O estado da arte em visão computacional atualmente é avaliado por *datasets*. Alguns *datasets* são utilizados como referência para treinamento e validação de modelos de redes neurais profundas entre eles é possível encontrar o CIFAR-10, CIFAR-100 (KRIZHEVSKY, 2014) e ImageNet (DENG, 2009).

Desde o começo da utilização de redes profundas na classificação de imagens a acurácia dos modelos tem aumentado. Em partes, esse fato se dá impulsionado pelo advento da internet, que permitiu que diversas pessoas, do mundo todo, pudessem contribuir para criar *datasets* com imagens rotuladas e agrupadas em classes. O trabalho em equipe dos diversos pesquisadores por todo o mundo impulsionou o campo de visão computacional para o patamar que alcançamos hoje (HABER, 2018).

O resultado desse esforço em conjunto gerou os *datasets* que utilizamos como *benchmark* para treinamento dos modelos de aprendizado. No começo os resultados eram expressivos, mas um pouco distante do estado da arte atual. Desde 2016 a acurácia tem crescido, juntamente com o tamanho e qualidade dos *datasets* mundiais (HUANG, 2018). No princípio da utilização de aprendizado profundo em visão computacional os primeiros valores de acurácia alcançavam cerca de 40 - 75% (SZEGEDY, 2016), (SZEGEDY, 2015), ao longo dos anos o amadurecimento da comunidade de visão computacional os levou aos, incríveis, 99% de acurácia em grandes datasets (HUANG, 2018), (KOWSARI, 2018).

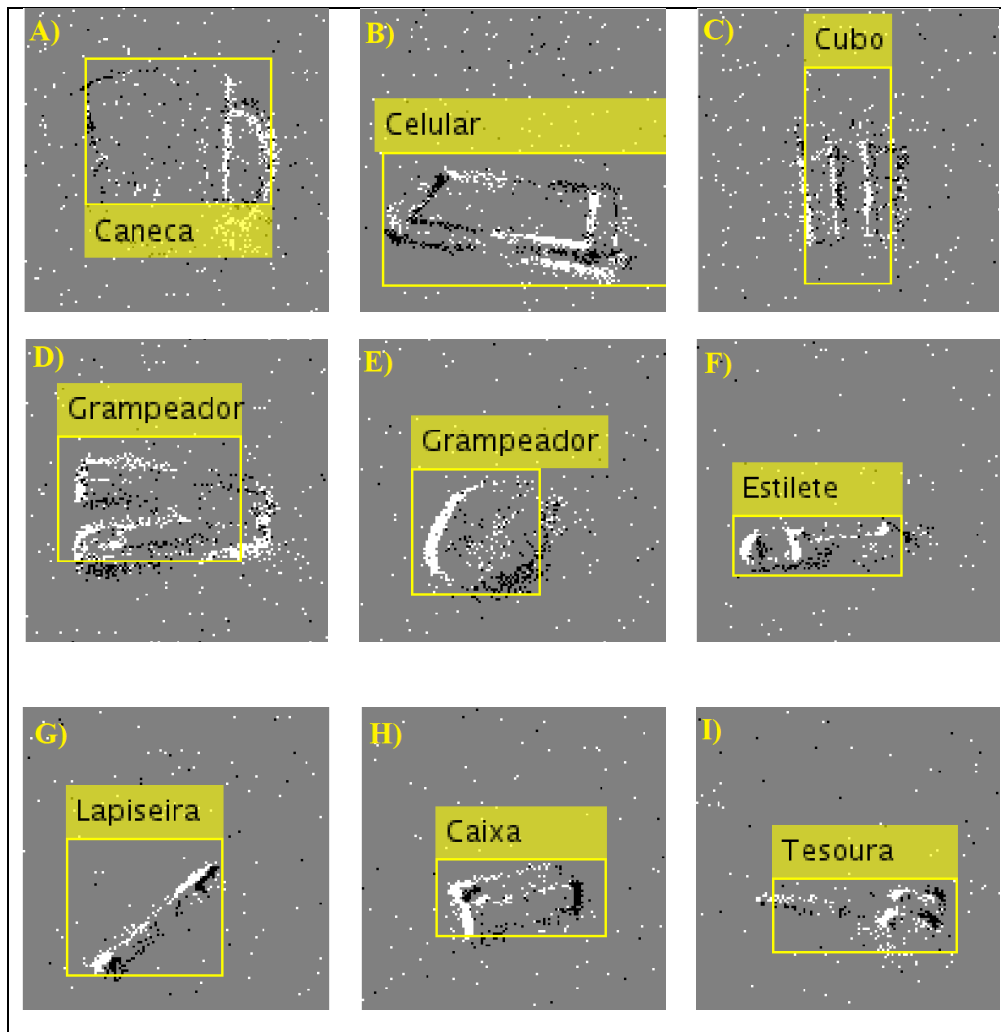
O engajamento científico na produção de tecnologia neuromórfica ganhou espaço a pouco tempo, porém já é possível visualizar movimentos de diversos cantos do mundo para produzir material neuromórfico (FRANCO, 2013), (INTEL, 2017), (LIU, 2015). Entretanto esse movimento

ainda é tímido, o que faz o campo de visão computacional neuromórfico ainda ser pequeno e pouco colaborativo, apesar de ser possível encontrar pesquisadores engajados em criar um *dataset* de imagens neuromórficas mesmo que ainda seja um dataset pouco maduro (TAN, 2018), (ORCHARD, 2015).

Resultados da aplicação de técnicas de aprendizado profundo em objetos gravados através de visão neuromórfica tem sido relatado nos últimos anos, o trabalho (GOPALAKRISHNAN, 2018) demonstra a aplicação de uma CNN no *dataset* Neuromorphic Caltech-101 (fruto do trabalho (ORCHARD, 2015)). A acurácia alcançada pelo modelo de aprendizagem profunda desenvolvido em (GOPALAKRISHNAN, 2018) alcançou um valor de 78.01%.

Outro trabalho similar (GHOSH, 2014) chegou a alcançar 99.10% de acurácia utilizando seus respectivos modelos de redes neurais convolutivas. Porém, a grande disparidade entre os resultados se dá pela falta de *datasets* para serem utilizados como *benchmark neuromórfico*. No trabalho (GHOSH, 2014) o *dataset* utilizado possui 8 classes de objetos e foi gravado utilizando objetos convencionais, enquanto o *dataset* utilizado por (GOPALAKRISHNAN, 2018) consiste em uma conversão das imagens do Caltech-101 gravados pelo sensor de visão neuromórfico posicionado em frente a uma tela de computador onde o conteúdo do Caltech-101 era apresentado ao sensor que realizava movimentos verticais em frente ao monitor para gerar os eventos.

Figura 15: Resultado da identificação e classificação do objeto pelo modelo de aprendizado profundo desenvolvido no trabalho. É possível perceber que o modelo se engana ao classificar o mouse (e) como um grampeador e, apesar de conseguir identificar o objeto na cena, por vezes, ele não consegue determinar todo o espaço ocupado pelo objeto (c)(a)(d).



Por fim, o modelo desenvolvido foi capaz de identificar o objeto na tela por meio de um processo de transferência de aprendizado que utilizamos para atribuir ao modelo de classificação desenvolvido a capacidade de determinar regiões. Esse modelo é conhecido como RPN (*Region Proposals Network*) e ele é composto por uma rede que determina a região de interesse (posição do objeto na cena) e então classifica o objeto achado (HE, 2017). Na Figura 15 é possível visualizar o modelo classificando e identificando as nove classes utilizadas nesse trabalho.

Alguns dos motivos para que a diferença entre a acurácia do trabalho desenvolvido por (GOPALAKRISHNAN, 2018) e a do modelo desenvolvido nesse trabalho pode ser consequência da utilização de uma câmera neuromórfica mais avançada, ou mesmo do ambiente de gravação dos objetos, outro fator que pode ser atribuído é a arquitetura da rede criada. A quantidade de atributos extraídos de cada classe, fruto da quantidade de camadas convolutivas, ou a qualidade do *dataset* utilizado ou mesmo a escolha dos objetos pertencentes a cada classe.

A escolha dos objetos também interfere na acurácia da rede. A heterogeneidade entre o formato dos objetos utilizadas no trabalho (GOPALAKRISHNAN, 2018) pode ser uma sugestão para a maior assertividade. Vide que o desempenho, da classificação de um celular pelos dois modelos, é semelhante; sendo 98.6% de acurácia no trabalho (GOPALAKRISHNAN, 2018) e o modelo proposto nesse trabalho alcançou 94.6%. O que demonstra a influência da escolha dos objetos para compor o *dataset* e ainda confirma a necessidade da criação de um *dataset* que sirva de *benchmark* para os trabalhos de visão neuromórfica.

5. Conclusões

Neste trabalho foi apresentado a implementação de uma rede neural convolutiva classificando imagens gravadas através de um sensor de visão neuromórfico. Foram utilizados métodos bioinspirados, tanto na aquisição da imagem, quanto no processamento da informação adquirida, esse processo resultou em um modelo de aprendizado profundo com acurácia geral de 75.31%. O que se mostra um resultado promissor frente aos desafios de se consolidar um campo de visão neuromórfico, que ainda se encontra em estágios iniciais.

Além da classificação dos objetos, o modelo desenvolvido também foi capaz de identificar a localização do objeto na cena, aplicando uma sinalização ao redor da posição do objeto. A implementação desse atributo foi alcançada através de um processo de transferência de aprendizado que complementou a capacidade do modelo, que além de classificar o objeto, passou a estimar a região ocupada pelo objeto na cena.

Porém, como o campo de visão neuromórfico ainda é relativamente jovem, quando comparado com visão computacional tradicional existem espaços não padronizados para desenvolvimento de diversas soluções pouco comparáveis entre si. Pois, apesar de ser vantajoso a heterogeneidade de estudos com os mais diversos propósitos e inspirações nos estágios iniciais do desenvolvimento de tal campo, cria-se um ambiente onde a validação de métodos satisfatórios se torna abstrato, visto que existe diferenças entre os objetos e a própria qualidade da informação dentro de cada *dataset*.

Contudo, um coproduto desse trabalho foi o começo da criação de um *dataset* neuromórfico, que estará disponível para a comunidade científica que tem interesse em desenvolver trabalhos nesse campo de pesquisa. De uma maneira colaborativa, é esperado que esse *dataset* inicial cresça e inspire outros trabalhos.

Aos horizontes futuros pretende-se explorar a quarta geração de redes neurais profundas, que se aproxima, cada vez mais, de sistemas bioinspirados ao utilizar padrões espaço-temporais dos *spikes* no tempo para classificar as informações.

6. Referências

SMITH, Christopher Upham Murray. **Biology of sensory systems**. John Wiley & Sons, 2008.

FRANCESCHINI, Nicolas; PICHON, Jean-Marc; BLANES, Christian. From insect vision to robot vision. **Phil. Trans. R. Soc. Lond. B**, v. 337, n. 1281, p. 283-294, 1992.

LIU, Shih-Chii et al. (Ed.). **Event-based neuromorphic systems**. John Wiley & Sons, 2015.

LIU, Shih-Chii; DELBRUCK, Tobi. Neuromorphic sensory systems. **Current opinion in neurobiology**, v. 20, n. 3, p. 288-295, 2010.

BOAHEN, Kwabena. Neuromorphic microchips. **Scientific American**, v. 292, n. 5, p. 56-63, 2005.

DOUGLAS, Rodney; MAHOWALD, Misha; MEAD, Carver. Neuromorphic analogue VLSI. **Annual review of neuroscience**, v. 18, n. 1, p. 255-281, 1995.

DENG, Jia et al. Imagenet: A large-scale hierarchical image database. In: **Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on**. Ieee, 2009. p. 248-255.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. 2012. p. 1097-1105.

JIMENEZ-FERNANDEZ, Angel et al. Neuro-inspired system for real-time vision sensor tilt correction. In: **Proceedings of 2010 IEEE International Symposium on Circuits and Systems**. IEEE, 2010. p. 1394-1397.

BARUA, Souptik; MIYATANI, Yoshitaka; VEERARAGHAVAN, Ashok. Direct face detection and video reconstruction from event cameras. In: **2016 IEEE winter conference on applications of computer vision (WACV)**. IEEE, 2016. p. 1-9.

KUFFLER, Stephen W. Discharge patterns and functional organization of mammalian retina. **Journal of neurophysiology**, v. 16, n. 1, p. 37-68, 1953.

MASLAND, Richard H. The fundamental plan of the retina. **Nature neuroscience**, v. 4, n. 9, p. 877, 2001.

RODIECK, R. W. The primate retina. **Comp Primate Biol**, v. 4, p. 203-278, 1988.

WERBLIN, Frank S.; DOWLING, John E. Organization of the retina of the mudpuppy, *Necturus maculosus*. II. Intracellular recording. **Journal of neurophysiology**, v. 32, n. 3, p. 339-355, 1969.

HALL, John E. **Guyton E Hall Tratado De Fisiologia Médica**. Elsevier Brasil, 2011, 12ed, p. 631-687.

SHARPE, L. T. et al. Temporal and spatial summation in the human rod visual system. **The Journal of Physiology**, v. 463, n. 1, p. 325-348, 1993.

BARLOW, Horace B. Summation and inhibition in the frog's retina. **The Journal of physiology**, v. 119, n. 1, p. 69-88, 1953.

SCHADE, Otto H. Optical and photoelectric analog of the eye. **JoSA**, v. 46, n. 9, p. 721-739, 1956.

ENROTH-CUGELL, Christina; ROBSON, John G. The contrast sensitivity of retinal ganglion cells of the cat. **The Journal of physiology**, v. 187, n. 3, p. 517-552, 1966.

Perkins, H. D. (22 de Junho de 2018). *Human eye*. Fonte: Encyclopædia Britannica : <https://www.britannica.com/science/human-eye/The-work-of-the-retina>

INTEL. The emergent capabilities in artificial intelligence being driven by Intel Labs have more in common with human cognition than with conventional computer logic. [S. /], 2017. Disponível em: <https://www.intel.com/content/www/us/en/research/neuromorphic-computing.html>. Acesso em: 2 jan. 2019.

BRANDLI, Christian et al. A 240× 180 130 db 3 μ s latency global shutter spatiotemporal vision sensor. **IEEE Journal of Solid-State Circuits**, v. 49, n. 10, p. 2333-2341, 2014.

POSCH, Christoph; MATOLIN, Daniel; WOHLGENANT, Rainer. An asynchronous time-based image sensor. In: **Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on**. IEEE, 2008. p. 2130-2133.

DELBRUCK, Tobi; LANG, Manuel. Robotic goalie with 3 ms reaction time at 4% CPU load using event-based dynamic vision sensor. **Frontiers in neuroscience**, v. 7, p. 223, 2013.

DRAZEN, David et al. Toward real-time particle tracking using an event-based dynamic vision sensor. **Experiments in Fluids**, v. 51, n. 5, p. 1465, 2011.

CENSI, Andrea et al. Low-latency localization by Active LED Markers tracking using a Dynamic Vision Sensor. In: **2013 IEEE/RSJ International Conference on Intelligent Robots and Systems**. IEEE, 2013. p. 891-898.

CHEN, Guang et al. Multi-cue event information fusion for pedestrian detection with neuromorphic vision sensors. **Frontiers in neurorobotics**, v. 13, p. 10, 2019.

ANDREOPOULOS, Alexander et al. A Low Power, High Throughput, Fully Event-Based Stereo System. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2018. p. 7532-7542.

GRIFFIN, Gregory; HOLUB, Alex; PERONA, Pietro. Caltech-256 object category dataset. 2007.

LIANG, Junwei et al. Peeking into the Future: Predicting Future Person Activities and Locations in Videos. **arXiv preprint arXiv:1902.03748**, 2019.

RINALDO, Francesca et al. A Computer Vision System to Detect Bedside Patient Mobilization. 2019.

SIMON, Martin et al. Complex-YOLO: An Euler-Region-Proposal for Real-Time 3D Object Detection on Point Clouds. In: **European Conference on Computer Vision**. Springer, Cham, 2018. p. 197-209.

SHAFIEE, Mohammad Javad et al. Fast YOLO: A fast you only look once system for real-time embedded object detection in video. **arXiv preprint arXiv:1709.05943**, 2017.

CORNFIELD, JEROME. Statistical classification methods. In: Proceedings of the Second Conference on the Diagnostic Process, Computer Diagnosis and Diagnostic Methods, Chicago. 1972. p. 108-130.

DEVIJVER, Pierre A.; KITTLER, Josef. **Pattern recognition: A statistical approach**. Prentice hall, 1982.

FUKUNAGA, Keinosuke. Introduction to statistical pattern recognition. Elsevier, 2013.

WENG, Juyang; AHUJA, Narendra; HUANG, Thomas S. Cresceptron: a self-organizing neural network which grows adaptively. In: **[Proceedings 1992] IJCNN International Joint Conference on Neural Networks**. IEEE, 1992. p. 576-581.

HUBEL, David H. **Eye, brain, and vision**. New York: Scientific American Library, 1988.

HUBEL, David H.; WIESEL, Torsten N. Brain mechanisms of vision. **Scientific American**, v. 241, n. 3, p. 150-163, 1979.

GEHRIG, Daniel et al. Asynchronous, photometric feature tracking using events and frames. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. 2018. p. 750-765.

POSCH, Christoph; MATOLIN, Daniel; WOHLGENANNT, Rainer. An asynchronous time-based image sensor. In: **2008 IEEE International Symposium on Circuits and Systems**. IEEE, 2008. p. 2130-2133.

CARPENTER, Roger HS. **Movements of the Eyes, 2nd Rev.** Pion Limited, 1988.

KUO, C.-C. Jay. Understanding convolutional neural networks with a mathematical model. **Journal of Visual Communication and Image Representation**, v. 41, p. 406-413, 2016.

HE, Kaiming et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: **Proceedings of the IEEE international conference on computer vision**. 2015. p. 1026-1034.

MURRAY, Naila; PERRONNIN, Florent. Generalized max pooling. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. 2014. p. 2473-2480.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. 2012. p. 1097-1105.

YADAV, Sanjay; SHUKLA, Sanyam. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: **2016 IEEE 6th International Conference on Advanced Computing (IACC)**. IEEE, 2016. p. 78-83.

FRANCO, Jesús Armando García; DEL VALLE PADILLA, Juan Luis; CISNEROS, Susana Ortega. Event-based image processing using a neuromorphic vision sensor. In: **2013 IEEE International Autumn Meeting on Power Electronics and Computing (ROPEC)**. IEEE, 2013. p. 1-6.

KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **Ijcai**. 1995. p. 1137-1145.

SHAO, Jun. Linear model selection by cross-validation. **Journal of the American statistical Association**, v. 88, n. 422, p. 486-494, 1993.

KRIZHEVSKY, Alex; NAIR, Vinod; HINTON, Geoffrey. The CIFAR dataset. online: <http://www.cs.toronto.edu/kriz/cifar.html>, v. 55, 2014.

DENG, Jia et al. Imagenet: A large-scale hierarchical image database. In: **2009 IEEE conference on computer vision and pattern recognition**. IEEE, 2009. p. 248-255.

HUANG, Yanping et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. **arXiv preprint arXiv:1811.06965**, 2018.

TAN, Cheston; LALLEE, Stephane; ORCHARD, Garrick. Benchmarking neuromorphic vision: lessons learnt from computer vision. **Frontiers in neuroscience**, v. 9, p. 374, 2015.

HABER, Nick et al. Modeling the scientist in the crib. **Journal of Vision**, v. 18, n. 10, p. 10-10, 2018.

ORCHARD, Garrick et al. Converting static image datasets to spiking neuromorphic datasets using saccades. **Frontiers in neuroscience**, v. 9, p. 437, 2015.

SZEGEDY, Christian et al. Rethinking the inception architecture for computer vision. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2016. p. 2818-2826.

SZEGEDY, Christian et al. Going deeper with convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. 2015. p. 1-9.

KOWSARI, Kamran et al. Rmdl: Random multimodel deep learning for classification. In: **Proceedings of the 2nd International Conference on Information System and Data Mining**. ACM, 2018. p. 19-28.

GHOSH, Rohan et al. Real-time object recognition and orientation estimation using an event-based camera and CNN. In: **2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings**. IEEE, 2014. p. 544-547.

GOPALAKRISHNAN, Roshan; CHUA, Yansong; IYER, Laxmi R. Classifying neuromorphic data using a deep learning framework for image classification. In: **2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)**. IEEE, 2018. p. 1520-1524.

HE, Kaiming et al. Mask r-cnn. In: **Proceedings of the IEEE international conference on computer vision**. 2017. p. 2961-2969.