SISTEMA ESTIMADOR DE NÚMERO DE PESSOAS EM RECINTOS POR ANÁLISE DE IMAGENS

Juliana Nunes da Silva
Graduanda em Engenharia Eletrônica
campus São Paulo - IFSP

Pamela Klebis Nogueira
Graduanda em Engenharia Eletrônica
campus São Paulo - IFSP

Ricardo Pires
Doutor em Sistemas Automáticos e Microeletrônicos
Université Montpellier II
docente no campus São Paulo - IFSP

Resumo

Este trabalho tem como proposta desenvolver um sistema capaz de estimar o número de pessoas dentro de um recinto a partir do uso de diferentes técnicas, tais como YOLO, LBPH, rede neural perceptron multicamadas e interpolação polinomial com diferenças entre imagens em relação à do recinto vazio. Nos experimentos, foram usadas imagens sintéticas criadas por meio do programa Blender. Com o YOLO, observou-se uma boa acurácia apenas em ambientes com poucos indivíduos, com sua contagem apresentando saturação em 16 pessoas, devido ao efeito de oclusão para números maiores. LBPH apresentou seus melhores resultados na faixa de 40 a 90 pessoas, com erro absoluto menor do que 20 pessoas nessa faixa. A interpolação polinomial com diferenças entre imagens apresentou melhorias na acurácia da contagem em relação às outras técnicas, com erro relativo máximo de 13% nas estimativas para números de pessoas a partir de 70. A aplicação de redes neurais perceptron multicamadas buscando-se combinar os resultados anteriores não apresentou resultados satisfatórios.

Palavras-chave: YOLO; LBPH; rede neural; processamento de imagens; contagem de pessoas.

SYSTEM FOR ESTIMATING THE NUMBER OF PEOPLE IN ENCLOSURES BY IMAGE ANALYSIS

Abstract

This work aims to develop a system capable of estimating the number of people inside a room using different techniques, such as YOLO, LBPH, multilayer perceptron neural network and polynomial interpolation with differences between images in relation to the empty room. In the experiments, synthetic images created using the Blender program were used. With YOLO, good accuracy was observed only in environments with few individuals, with its count showing saturation at 16 people, due to the occlusion effect for larger numbers. LBPH presented its best results in the range of 40 to 90 people, with an absolute error of less than 20 people in this range. Polynomial interpolation with differences between images showed improvements in counting accuracy in relation to other techniques, with a maximum relative error of 13% in estimates for numbers of people from 70 onwards. The

application of multilayer perceptron neural networks seeking to combine the previous results did not present satisfactory results.

Keywords: YOLO; LBPH; neural network; image processing; people counting.

Introdução

Os grandes centros urbanos possuem altas concentrações populacionais. A cidade de São Paulo (SP) conta com aproximadamente 11,5 milhões de habitantes, de acordo com o último censo divulgado (IBGE 2022). As aglomerações são cenas comuns nesse ambiente e trazem diversos desafios na elaboração de soluções para problemas advindos do grande número de habitantes, como problemas de infraestrutura, superlotação dos transportes em horários de pico, trânsito e acesso dessa população aos serviços essenciais, como saúde, educação e trabalho. Em 2008, o metrô de São Paulo foi considerado mais lotado do que o de Tóquio, no Japão (Folha de São Paulo, 2008). Para elaborar políticas eficientes na solução desses problemas, é fundamental ter os números aproximados de usuários que utilizam esses serviços. Para se fazer isso com uma população tão grande quanto a de São Paulo, por exemplo, a tecnologia deverá ser usada na busca de soluções que ajudem no levantamento desses dados, como estimar quantas pessoas utilizam o metrô diariamente, a fim de otimizar esse serviço para o bem estar da população.

A superlotação também acomete diversas outras cidades do Brasil, além de São Paulo. Em 2013, ocorreu um incêndio na Boate Kiss, em Santa Maria, no Rio Grande do Sul. A casa noturna era capaz de comportar por volta de 770 pessoas, mas estima-se que mais de mil participantes estavam no recinto. O excesso de pessoas foi um dos motivos que contribuíram para a tragédia (G1, 2015). Eventos esportivos, festivais e shows com saturação de público também colocam em risco a segurança dos presentes. Em 2010, na Alemanha, ocorreu um acidente desse tipo que deixou 19 mortos, porque havia quase seis vezes mais pessoas do que era comportado no local (G1, 2010).

Nos presídios brasileiros, entre 2011 e 2021, estima-se que havia 66% mais de população carcerária do que era comportada (CNJ, 2022). Situações como essas são comuns e causam prejuízos ao espaço e às pessoas - más condições de saúde e higiene, falta de privacidade, mortes, acidentes e insatisfação.

Como visto, o Brasil é um país que possui muitos problemas com o acúmulo de pessoas em ambientes. Dessa forma, se faz necessário estimar o número de pessoas em recintos, como medida de segurança para se reduzir o número de acidentes e tumultos, permitir o planejamento

de itens de seguridade necessários para evacuações, prevenir contaminações, como durante a pandemia do coronavírus em que o distanciamento social foi essencial. Além disso, também auxilia na coleta de resíduos orgânicos por permitir planejar com antecedência a quantidade de banheiros químicos necessários durante um evento, por exemplo, e com isso diminuir os impactos ambientais. Por todos esses aspectos supracitados, um estimador de número de pessoas seria de grande utilidade.

No Brasil, quando ocorrem grandes aglomerações ao ar livre, em certos eventos, estimativas do número de pessoas presentes costumam ser divulgadas pela Polícia Militar e por institutos de pesquisa. Em suas estimativas, o Datafolha leva em conta a área ocupada e a densidade aproximada de pessoas naquela área. A Polícia Militar de São Paulo usa imagens aéreas, obtidas a partir de um helicóptero, para calcular a área ocupada e assume uma determinada densidade de ocupação, como cinco pessoas por metro quadrado (Veja São Paulo, 2016).

O trabalho aqui desenvolvido é restrito à estimativa do número de pessoas em um recinto fechado, a partir de imagens obtidas por uma única câmera em posição fixa. Esta é uma configuração útil e viável em situações práticas. A estimativa do número total de pessoas no recinto deverá ser feita a partir de imagens de apenas a parte do recinto que esteja no campo de visão da câmera, ou seja, sem que necessariamente todas as pessoas presentes apareçam nas imagens.

Objetivos

O objetivo geral deste trabalho é o projeto de um sistema estimador de número de pessoas num recinto, usando processamento de imagens associado a aprendizagem de máquina e outras técnicas. Os objetivos específicos são:

- Formar um banco de imagens adequado.
- Aplicar várias técnicas, para compará-las e identificar a melhor dentre elas.
- Boas estimativas dos números totais de pessoas deverão ser obtidas, mesmo que as imagens capturadas sejam parciais, não cobrindo toda a área do recinto e, consequentemente, não registrando todas as pessoas presentes.

Estudo Bibliográfico

Um estudo bibliográfico sobre o tema deste trabalho foi realizado utilizando-se o sistema de buscas Google Acadêmico. As palavras-chave usadas foram, em conjunto: *review survey crowd images counting estimation number people*. Dentre os muitos artigos resultantes da busca pelo sistema, foram selecionados para análise pelos autores, com base em seus títulos e resumos, os artigos comentados a seguir.

O artigo de Sindagi e Patel (2018) traz um levantamento do estado da arte do uso de redes neurais convolucionais (RNC) na contagem de pessoas em multidões a partir de uma imagem única (não vídeo). Os autores justificam o foco do levantamento nos trabalhos que usaram RNC alegando que essa técnica tem demonstrado melhorias em relação a técnicas anteriores, as quais dependiam de processos manuais de extração de características das imagens para processamento. Apesar do foco no uso de RNC, os autores afirmam que, dentre os vários trabalhos que usaram outras técnicas, alguns usaram regressão, buscando aprender um mapeamento entre características extraídas de trechos de cada imagem com seu número de pessoas. Quanto ao uso de RNC, os autores defendem a sua superioridade, mas apontam o fato de que o treinamento daquelas redes requer a disponibilidade de enormes bancos de imagens rotuladas manualmente com o número de pessoas visíveis em cada uma delas. As técnicas baseadas em RNC foram subdivididas entre aquelas que analisam cada imagem por inteiro e aquelas que a analisam por partes, usando uma janela deslizante sobre ela. Os trabalhos comentados naquele artigo usaram bancos de dados variados. Os melhores resultados lá apresentados tiveram erro absoluto médio da ordem de uma ou duas unidades, obtidos em dois dos bancos de imagens (chamados UCSD e Mall).

Ilyas, Shahzad e Kim (2019) também fizeram uma revisão da literatura sobre o uso de redes neurais convolucionais (RNC) na contagem de pessoas em multidões. Eles mencionaram a existência, também, de outras técnicas usadas com essa finalidade, tais como: regressão, contagem por detecção, por estimativa de densidade e agrupamentos. Estas outras técnin cas apresentam alguns desafios, como questões de iluminação, variação da densidade, desorganização dos objetos e oclusões no espaço, que aumentam o erro de previsão e diminuem a acurácia. Para o caso específico do uso de RNC, foco daquela revisão, foram mencionadas como dificuldades que degradam a acurácia: a oclusão (objetos muito próximos um ao outro ou um diante de outro), não uniformidade na distribuição dos objetos na cena, não uniformidade na escala dos objetos na cena e não uniformidade na perspectiva, devido às diferentes posições

dos objetos em relação à câmera. Os autores alegam que as RNC lidam melhor do que outras técnicas com essas dificuldades. Porém, ressaltam que, comparativamente, as RNC consistem numa arquitetura mais complexa, com número maior de parâmetros a serem ajustados, maior custo computacional e maior dificuldade em conseguir operar em tempo real. Assim como Sindagi e Patel (2018), também foi discutida a possibilidade de se analisar as imagens por partes.

Em outro estudo da literatura em contagem automática de pessoas em multidões, em muitos aspectos similar aos comentados anteriormente, Gouiaa, Akhloufi e Shahbazi (2021) inovaram ao incluir uma análise da extensão das técnicas de contagem de pessoas à contagem de outros tipos de objetos e ao incluir trabalhos mais recentes, em que as imagens são obtidas por meio de drones ou de outras aeronaves. Assim como em Sindagi e Patel (2018), os melhores resultados lá apresentados tiveram erro absoluto médio da ordem de uma ou duas unidades, para algumas combinações de técnicas e de bancos de imagens.

Na revisão da literatura de Li et al (2021) sobre abordagens na contagem automática de multidões, os autores analisam várias das técnicas empregadas e também concluem que a tendência é a de se adotarem, cada vez mais, técnicas que incluam a chamada aprendizagem profunda, ou seja, técnicas baseadas em RNC com muitas camadas. Muitos dos aspectos dessa revisão são similares aos das revisões comentadas nos parágrafos anteriores.

Segundo Moraes (2021), muitos estudos vêm sendo direcionados para a área de contagem de pessoas, devido a sua aplicação para gerar dados mais completos no que diz respeito a multidões e seu comportamento. Ele destaca ainda as transformações das redes neurais que vêm sendo utilizadas no desenvolvimento dessa tecnologia. Uma delas é o uso de aprendizagem profunda. No foco daquela pesquisa - contar o fluxo de clientes em um estabelecimento - utilizou-se o programa YOLOv5 (The Linux Foundation, 2023). Ao final, pode-se entender que tanto a altura em que é posicionada a câmera que fará a captura de imagem quanto a angulação dos corpos das pessoas influenciam nos resultados. O YOLOv5 obteve bons resultados com alta acurácia e alta velocidade ao processar os dados.

Deve-se observar que os trabalhos encontrados na literatura buscaram contar apenas as pessoas visíveis em cada imagem, diferentemente do objetivo deste trabalho, que é o de estimar o número total de pessoas presentes num recinto, incluindo as que estiverem fora do campo de visão da câmera. Isto requer que o banco de imagens usado aqui tenha cada imagem rotulada com o número de pessoas que se sabe estarem no recinto em cada caso, não apenas com as pessoas visíveis.

Fundamentação Teórica

É comum representar-se uma imagem no formato digital como uma ou mais matrizes. Cada posição na matriz contém um *pixel* (elemento da imagem, de *picture element*). Na representação em tons de cinza, cada elemento da matriz contém um número proporcional à claridade da imagem naquele ponto. Uma forma comum de representação de imagens em cores é o sistema RGB (vermelho, verde e azul, de *red*, *green*, *blue*). Nele, uma imagem é representada por três matrizes, uma matriz para a intensidade de cada uma dessas componentes de cor (Gonzalez; Woods, 2017). Renderizar uma imagem é o processo de se converter símbolos gráficos usados por um programa de modelagem bidimensional ou tridimensional num arquivo visual, com pormenores, efeitos de iluminação e realismo (Tewari et al, 2020).

Uma rede neural artificial é uma estrutura inspirada no cérebro dos animais superiores. Ela possui neurônios interconectados, possivelmente, em várias camadas. Cada neurônio possui várias entradas, às quais são aplicados números. Ao entrar no neurônio, cada um desses números é multiplicado por um coeficiente, chamado de peso, a ser ajustado durante uma fase de aprendizado. A soma destes produtos é aplicada como entrada a uma função não linear, chamada de função de ativação, a qual calcula o valor de saída do neurônio. Este valor de saída, por sua vez, pode ser usado como um dos valores de entrada de um neurônio de uma próxima camada ou pode ser um valor de saída da rede neural. Durante uma fase de aprendizado (ou de treino), muitos dados conhecidos são aplicados às entradas da rede. Para cada um desses dados, a rede calcula um valor de saída. Este é comparado com o valor desejado para aquele dado. Caso haja diferença (erro) entre esses valores, ocorre um processo de retropropagação, pelo qual o valor do erro é usado para que sejam feitas correções nos valores atuais dos pesos da rede, de forma a se reduzir o valor daquele erro. Esse processo continua, com a reapresentação dos exemplos de treino, em várias ordens, até que os erros nas saídas fiquem muito pequenos ou nulos. Então, considera-se que a rede está treinada. Espera-se que, a partir desse momento, quando novos dados forem aplicados à entrada, a rede apresente, à saída, valores compatíveis com o que ela aprendeu na fase de treino. Assim, ela é capaz de aprender uma relação não linear entre vetores de entrada e valores de saída, os quais podem representar classes ou valores numéricos numa determinada faixa (Haykin, 2007). Ela pode ser utilizada em diversas aplicações, tais como controle de qualidade, diagnósticos médicos, predições financeiras, etc. (Amazon, 2023).

A Rede Neural Convolucional (CNN, de *Convolutional Neural Network*) é um avanço das redes neurais de múltiplas camadas inspirada no processo biológico de processamento de dados visuais. A CNN tem como uma de suas características a presença de filtros em dados visuais para imagens, fazendo com que se leve em conta a vizinhança entre os *pixels* no processamento realizado pela rede. Suas principais aplicações são nas áreas de classificação, detecção e reconhecimento de imagens ou vídeos. A CNN possui camadas de convolução que são compostas por diversos neurônios artificiais e cada neurônio é capaz de adotar filtros em partes específicas da imagem a ser tratada. A rede utiliza neurônios dedicados a um conjunto de *pixels* da imagem a ser tratada. Esse processo de filtragem é feito para extrair características da imagem e entregá-las a uma rede neural (Vargas; Paes; Vasconcelos, 2016).

Dentre as diversas opções de CNN disponíveis a serem utilizadas para tratamento de imagens, uma comumente utilizada é o YOLO *(You Only Look Once)* (The Linux Foundation, 2023). Essa rede neural "olha" apenas uma vez para a imagem e a divide em pequenos blocos. Um processo probabilístico infere onde estão os objetos, analisando os blocos.

No domínio de análise de imagens, também é usada a técnica Histograma de Padrões Binários Locais (LBPH, de *Local Binary Pattern Histogram*). Ela é baseada em um operador binário. A representação obtida pelo LBP é calculada deslizando-se sobre a imagem de entrada uma matriz a qual, normalmente, tem formato 3x3. A imagem de entrada precisa estar em tons de cinza. O *pixel* na imagem de entrada sob a posição central da matriz deslizante é comparado com os seus oito vizinhos. Para cada um dos vizinhos, é anotado um bit 0 quando ele for mais escuro e um bit 1 quando ele for mais claro do que o *pixel* central. Então, esses oito bits são concatenados, formando um número. Uma imagem resultante será formada com esses números, nas posições correspondentes às da imagem de entrada. Após esse processo, são calculados histogramas de ocorrências dos números na imagem resultante. Esses histogramas são associados à presença de determinadas texturas ou de determinada face na imagem de entrada (Galimberti, 2018). Para a classificação de uma nova imagem, deve-se repetir para ela o processo de formação dos histogramas. Estes, então, são comparados aos histogramas de imagens previamente analisadas. A nova imagem é classificada como pertencendo à classe das imagens que tenham histogramas mais parecidos com o dela. Uma das técnicas usadas para essa classificação é a k-NN (de *k-Nearest Neighbors*), a qual classifica um novo dados com base nas classes dos *k* dados mais parecidos previamente rotulados (Taunk et al, 2019).

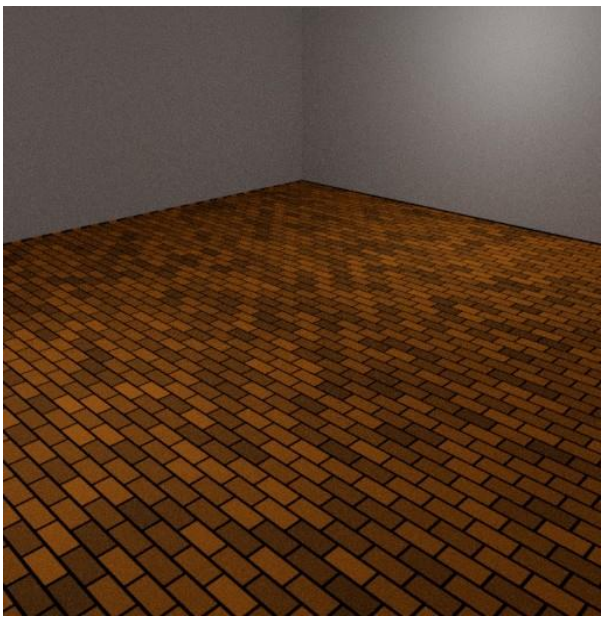
Em Matemática, a regressão é o processo de explicação do valor de uma variável observada a partir dos valores de outras variáveis de um problema. Para isso, a regressão busca

a construção de um modelo, o qual pode consistir numa expressão algébrica relacionando aquelas variáveis (Freund; Wilson; Sa, 2006).

Procedimento Experimental

Inicialmente, foi feita a renderização de imagens com pessoas dentro de um recinto. Devido à dificuldade em se encontrarem muitas imagens de um mesmo recinto com números de pessoas variados e conhecidos, além da possível necessidade de autorização para uso das imagens das pessoas, foi necessário utilizar um programa capaz de compor imagens sintéticas de pessoas em números variados em um espaço fechado. Foi escolhido o programa de uso livre e código aberto Blender (Blender Foundation, 2023). Com ele, é possível criar uma representação realista de um recinto, uma representação de figura humana e multiplicá-la em posições aleatórias no recinto na quantidade escolhida. No modelo de recinto criado (Figura 1), foi posicionada uma câmera numa certa posição acima da altura das pessoas, simulando-se uma situação real de monitoramento. O campo de visão da câmera não abrange o recinto completo. Algumas pessoas presentes no recinto poderão estar fora desse campo de visão. A resolução de cada imagem foi de 512x512 pixels. Não foi adotada resolução maior, devido ao longo tempo necessário para se renderizar cada imagem com o Blender num computador doméstico, tempo esse dependente do modelo do computador, tendo chegado à ordem de quinze minutos em computadores neste trabalho.

Figura 1: Recinto vazio, criado usando-se o programa Blender.



Fonte: Autores.

Foram empregadas 60 imagens especialmente criadas pelos autores para o processo de configuração do sistema, englobando variações na presença de homens e mulheres, composições exclusivamente femininas, bem como exclusivamente masculinas. A Figura 2 é um exemplo. O número de pessoas nas imagens variou de um mínimo de 10 a um máximo de 200 indivíduos. No procedimento, inicialmente, adotou-se modelo para YOLOv8 já treinado para detectar pessoas.

O referido modelo demonstra eficácia na detecção de pessoas a partir de seus corpos completos. Contudo, diante de situações de sobreposição de imagens de pessoas, ele falha na detecção daquelas que estão com seus corpos obstruídos na imagem pelos corpos de outras pessoas, resultando em inacurácias nas previsões. Deve-se lembrar também que, devido à posição particular da câmera, algumas pessoas presentes no ambiente não são capturadas na imagem, o que impede sua detecção pelo YOLO.

Figura 2 - Recinto com 35 homens e 35 mulheres, criado usando-se o programa Blender.

Fonte: Autores.

Os resultados de contagem de pessoas detectadas pelo YOLO para todas as imagens são apresentados na Tabela 1. Nela, a coluna H informa o número de homens no recinto, a coluna M o número de mulheres, a coluna T a soma de homens e mulheres (total) e a coluna Y o número detectado na imagem pelo YOLO.

Tabela 1 – Resultados YOLO, LBPH e Diferenças.

H (Homem)	M (Mulher)	T (Total)	Y (YOLO)	L (LBPH)	D (Diferenças)	H (Homem)	M (Mulher)	T (Total)	Y (YOLO)	L (LBPH)	D (Diferenças)
0	0	0	0	13,6	0	50	50	100	7	73,08	36190
0	10	10	10	29,45	11442	0	110	110	12	106,38	28170
5	5	10	9	30,54	14100	55	55	110	8	78,13	37074
10	0	10	8	29,73	17382	110	0	110	8	111,65	40487
20	0	20	14	41,05	22973	120	0	120	9	61,57	40785
10	10	20	14	41,95	20037	0	120	120	14	111,65	29139
0	20	20	14	37,99	14733	60	60	120	7	90,24	38035
15	15	30	15	53,85	23895	130	0	130	8	58,05	41340
0	30	30	14	44,39	16804	65	65	130	8	91,63	38381
30	0	30	12	58,04	30339	0	130	130	15	111,68	29727
20	20	40	16	59,13	26138	0	140	140	12	106,14	30804
0	40	40	14	52,82	18972	70	70	140	8	90,07	38761
40	0	40	9	57,59	33785	140	0	140	8	57,89	41778
50	0	50	9	57,59	33785	0	150	150	13	127,5	31476

25	25	50	13	61,48	27625	150	0	150	8	57,71	42125
0	50	50	16	53,66	20466	75	75	150	8	91,44	39338
30	30	60	14	63,66	28936	0	160	160	13	123,12	31896
60	0	60	6	64,73	36118	80	80	160	6	92,46	40024
0	60	60	14	79,19	23461	160	0	160	8	54,71	42125
35	35	70	10	71,26	31692	170	0	170	9	64,99	42953
0	70	70	13	77,24	23919	0	170	170	12	125,24	32299
70	0	70	9	63,16	37060	85	85	170	8	92,31	40570
40	40	80	12	78,36	33445	180	0	180	6	62,39	43184
80	0	80	9	61,33	38147	0	180	180	12	118,12	32587
0	80	80	12	82,79	25645	90	90	180	8	95,36	40765
45	45	90	10	75,58	34132	0	190	190	11	132,39	33780
0	90	90	14	94,34	26537	95	95	190	9	99,88	40921
90	0	90	9	61,57	39362	190	0	190	5	61,33	43201
100	0	100	9	58,55	40167	0	200	200	11	136,17	34899
0	100	100	14	93,63	27406	100	100	200	11	95,79	41175
						200	0	200	5	60,59	43350

Fonte: Autores.

Analisando-se os resultados, levando-se em conta que nem todas as pessoas no recinto estão no campo de visão da câmera, observa-se que o desempenho do modelo é satisfatório quando há poucas pessoas. Mas, à medida que ocorre sobreposição entre imagens de pessoas (oclusão), verifica-se que a acurácia na estimativa do número de pessoas diminui. O número máximo de pessoas detectadas pelo YOLO foi 16, mesmo quando havia mais de 100 pessoas no recinto.

Com o intuito de tentar aprimorar as previsões, optou-se por incorporar a técnica LBPH. A implementação usada foi aquela da biblioteca OpenCV de visão computacional (OpenCV Team, 2023), em linguagem de programação Python. Para aplicar esse método, uma imagem de exemplo, com 512x512 *pixels*, com figuras de homens no recinto, foi convertida para tons de cinza e, posteriormente, subdividida em 16 linhas e 16 colunas, resultando 256 retângulos. De cada retângulo, foi obtido um histograma por LBPH. Em seguida, realizou-se uma avaliação manual, na qual se anotou o número de cabeças presentes em cada retângulo, associando-se esse número, como rótulo, ao histograma correspondente. Em situações em que apenas metade de uma cabeça era visível, atribuiu-se o valor de 0,5 cabeça. De maneira semelhante, ao identificar um terço de uma cabeça, registrou-se o valor correspondente, e assim por diante.

Para todas as outras imagens, repetiu-se o procedimento: conversão de cada uma delas para tons de cinza, divisão em 256 retângulos, obtenção do histograma LBPH de cada retângulo e, usando-se o algoritmo k-NN com $k = 3$, fez-se a identificação dos histogramas rotulados

previamente mais parecidos, para se atribuir um número de cabeças aos novos retângulos sob avaliação. Para cada imagem, a soma dos números de cabeças detectadas nos 256 retângulos foi usada como a estimativa do número de pessoas detectadas. Na faixa de 40 a 90 pessoas, observou-se uma maior proximidade da estimativa do número de pessoas por esse método com LBPH do que com YOLO entre os resultados obtidos e os valores reais, conforme a Tabela 1, coluna L, enquanto fora dessa faixa, a acurácia diminuiu. Esse efeito pode ser atribuído, em parte, à avaliação manual nos retângulos, fator que introduz subjetividade nos resultados. Outro problema pode ter sido a insuficiência na resolução das imagens, já que, após a subdivisão delas em 256 retângulos, cada um deles ficou com apenas 32x32 *pixels*, o que pode dificultar a identificação de padrões nele. A diversidade de gênero também emergiu como um possível influenciador desses resultados não ideais, já que a identificação de cabeças com características distintas pode apresentar desafios, já que a rotulagem de retângulos para LBPH quanto à presença de cabeças foi feita com uma imagem que continha homens e não mulheres. Uma consideração importante para aprimorar a metodologia com LBPH seria o uso de imagens de maior resolução na avaliação manual, a inclusão de imagens com mulheres no processo de rotulagem, associados a uma redução na divisão em quadrados, para obter contagens mais acuradas de cabeças por retângulo.

Em seguida, também se fez um cálculo da diferença entre a imagem do recinto vazio e as imagens do recinto com pessoas. Quanto maior a diferença, maior a discrepância entre as imagens causada pela presença de pessoas. Para isso, a OpenCV possui uma função chamada *norm*, a qual recebe, como parâmetros, duas imagens e calcula a norma da diferença entre elas. Essa norma consiste na raiz quadrada da soma dos quadrados das diferenças dos valores *pixel* a *pixel* entre as duas imagens. A diferença pode ser visualizada na Tabela 1, coluna D. Vê-se que, inicialmente, ela aumenta com o aumento no número de pessoas e que, para números maiores de pessoas, ela tende a se estabilizar. Isso ocorre porque, para pequenos números de pessoas, o acréscimo de uma pessoa na imagem tende a ocupar uma região que era igual à da imagem do recinto vazio, ao passo que, para grandes números de pessoas, o acréscimo de uma pessoa tende a ocupar na imagem uma região que já não era mais igual à do recinto vazio, por já estar ocupada por outras pessoas.

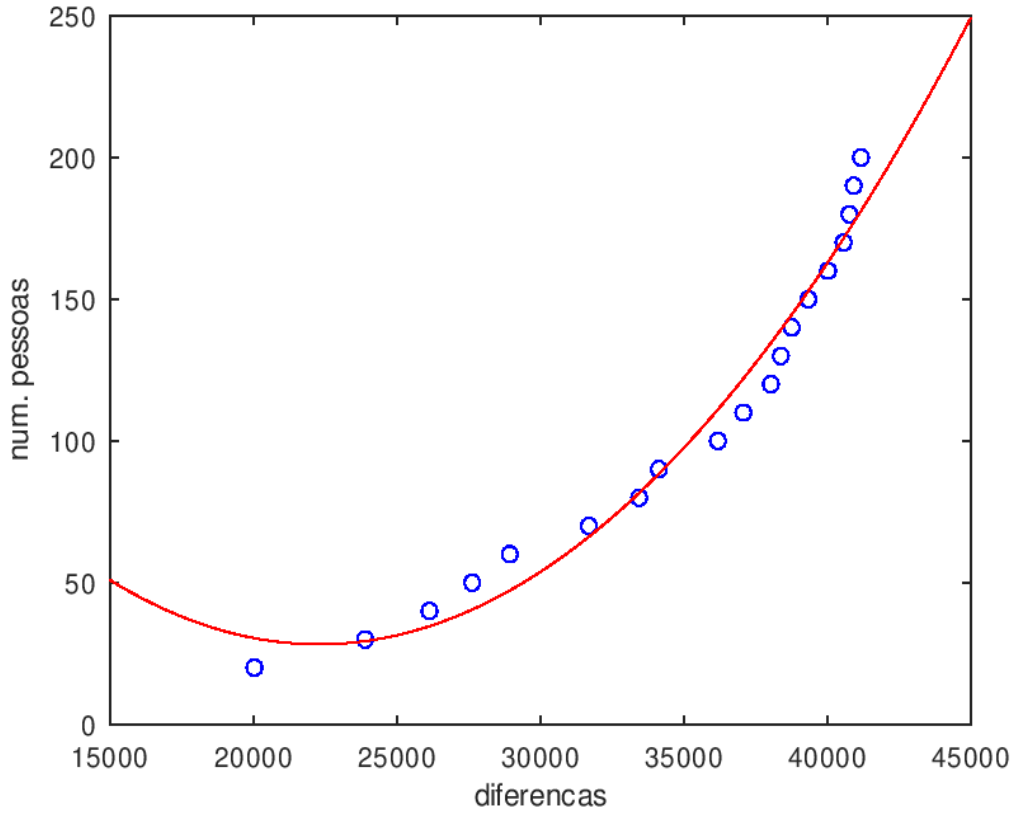
Tabela 2 - Resultados com a rede neural perceptron multicamadas.

Número de Neurônios Camada 1	Número de Neurônios Camada 2	Erro Absoluto Médio

1	-	101,08
5	-	99,28
10	-	97,40
15	-	98,14
20	-	94,04
30	-	93,73
40	-	86,25
50	-	83,99
100	-	67,05
200	-	43,48
300	-	27,29
1000	-	11,75
20	5	73,12
30	5	69,47
40	5	30,43
50	5	41,89

Fonte: Autores.

Figura 3: Gráfico da interpolação polinomial de grau 2 da relação entre o número de pessoas no recinto e a diferença da imagem em relação à do recinto vazio.



Fonte: Autores.

Buscando-se melhorar os resultados obtidos, criou-se uma rede neural perceptron multicamadas para relacionar, de forma combinada, os valores obtidos do YOLO, LBPH e da diferença com o valor total real de pessoas no recinto. Para isso, utilizou-se o pacote Keras (Keras, 2023) para executar uma rede neural de forma a obter uma codificação simples em linguagem Python. Foi feita uma normalização dos dados, tendo em vista que os valores da diferença entre imagens, quando comparados com os dos resultados do YOLO e do LBPH, eram muito grandes. Realizou-se treinamento com validação cruzada, com 80% dos dados para treino e 20% validação. Na validação cruzada, para várias configurações da rede neural (números de neurônios e números de camadas), não foi encontrada uma configuração com média de erro absoluto abaixo de 11,75 nas amostras de validação. Os valores obtidos são relativamente altos diante dos números totais de pessoas no recinto nos exemplos disponíveis, com exceção de quando se usam 1000 neurônios, caso em que se obteve a média de erro absoluto de 11,75.

Os resultados de diferenças entre imagens com pessoas e a imagem do recinto vazio foram utilizados, em seguida, para interpolação polinomial entre a coluna T (total) e a coluna

D (diferença) da Tabela 1. Observou-se que essa interpolação, para graus do polinômio 2 ou 3, apresentou um desempenho mais consistente em recintos com mais pessoas. Em cenários com poucas pessoas, os resultados não foram tão satisfatórios, sendo a maioria dos valores obtidos superior a 100. Usando-se, então, apenas as linhas da Tabela 1 com valores de diferença (coluna D) superiores a 20000 e os casos em que havia tanto homens quanto mulheres nas imagens, obteve-se, com o programa Octave (Eaton, 2023), um polinômio interpolador de grau 2 entre o valor da diferença entre imagens e o número total de pessoas no recinto. O resultado dessa interpolação é dado graficamente na Figura 3.

Tabela 3 - Resultados da validação do polinômio interpolador de grau 2 usando-se a diferença entre imagens.

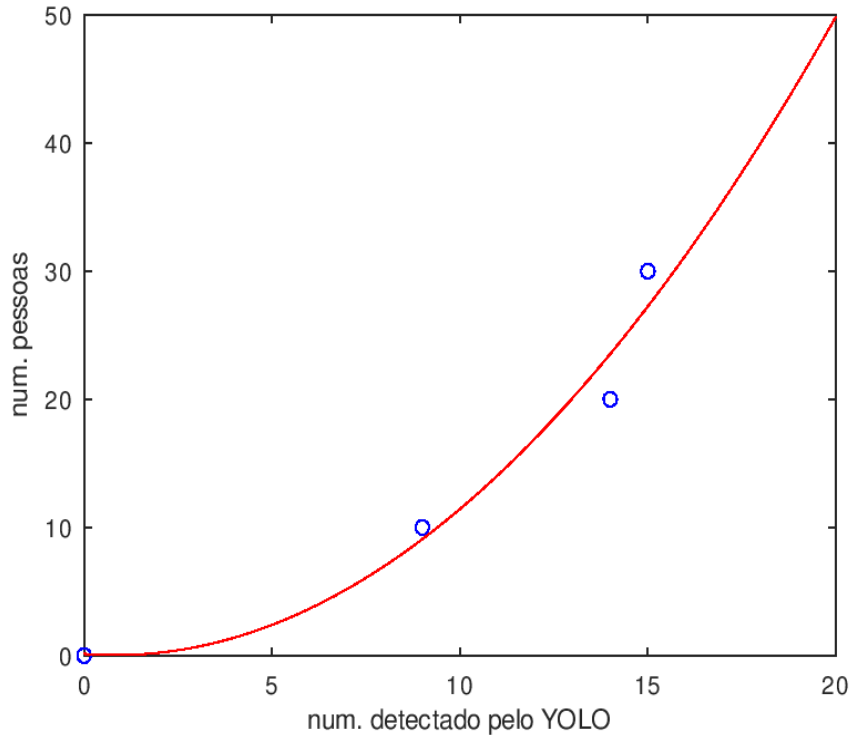
Número de pessoas do caso retirado	Número estimado pelo polinômio a partir dos demais casos	Erro (%)
30	29	-2,7
40	34	-16
50	39	-22
60	45	-25
70	66	-6,4
80	82	+2,4
90	88	-2,0
100	112	+12
110	123	+12
120	136	+13
130	140	+7,7
140	145	+3,5
150	153	+2,1
160	163	+2,2
170	172	+1,0
180	174	-3,5
190	175	-8,1

Fonte: Autores.

Para validação da interpolação polinomial, após a verificação inicial de sua viabilidade (pela Figura 3), ela foi realizada novamente repetidas vezes. Em cada vez, um dos casos (círculos na Figura 3) foi retirado do conjunto. Um polinômio interpolador de grau 2 foi

calculado para os demais casos. Então, esse polinômio foi usado para se estimar o número de pessoas do caso retirado. Os resultados estão na Tabela 3. Para esses resultados, o erro absoluto médio foi de 7,41, menor do que aqueles obtidos por rede neural perceptron multicamadas (Tabela 2).

Figura 4: Gráfico da interpolação polinomial de grau 2 da relação entre o número de pessoas no recinto e o número de pessoas detectadas pelo YOLO, para alguns casos do início da Tabela 1.



Fonte: Autores.

Em situações em que a diferença está abaixo de um limiar predeterminado, fixado em 20000, opta-se por utilizar a estimativa fornecida pelo YOLO. Para os primeiros casos da Tabela 1, incluindo o caso do recinto vazio e casos contendo tanto homens quanto mulheres, foi obtido o polinômio interpolador de grau 2, cujo gráfico ajustado aos valores de detecção pelo YOLO está na Figura 4.

Nessas situações, não foi feito um estudo do erro similar ao do caso da interpolação polinomial baseada na diferença entre imagens, porque ainda são exemplos pouco numerosos e de menor interesse neste trabalho, uma vez que o foco foi o de estimar com boa acurácia

números de pessoas no recinto em casos de lotação, sendo, portanto, de menor interesse os casos com poucas pessoas.

Conclusão

Neste trabalho, foi realizada uma revisão da literatura para avaliar uma variedade de abordagens e métodos para estimar pessoas em diferentes ambientes. Essas abordagens incluem exemplos com aprendizagem profunda (por exemplo, YOLO) e métodos mais tradicionais. O estudo encontrou problemas frequentes, como variação de densidade, de iluminação, obstruções e desorganização de objetos nas imagens.

A metodologia que se adotou para este trabalho englobou a criação de imagens sintéticas com o software Blender, a aplicação do YOLO para detecção de pessoas e introdução da técnica LBPH para que a acurácia da contagem de indivíduos fosse melhorada em situação de sobreposição. Com os resultados obtidos foi revelado que o YOLO apresentou uma boa acurácia para poucas pessoas dentro de um recinto. Em situações de aglomerações, ela diminuiu. Em contrapartida, o LBPH demonstrou boa acurácia na faixa de 40 a 90 pessoas.

A técnica de rede neural combinando os resultados das outras técnicas só foi capaz de obter resultados razoáveis quando o número de neurônios foi da ordem de milhares. Isso pode estar relacionado com a baixa qualidade das detecções do YOLO para números grandes de pessoas e com o fato de o LBPH só ter funcionado bem para uma faixa limitada de número de pessoas. A técnica de interpolação polinomial baseada nas diferenças entre imagens foi a que proporcionou a maior acurácia, quando usada a partir do valor de diferença de 20000 obtida pela função *norm* da OpenCV. Conforme visto na seção de resultados, os valores observados pelo YOLO podem ser aproveitados quando a diferença estiver menor do que 20000.

Pode-se admitir que os resultados aqui obtidos usando-se imagens sintéticas não correspondam totalmente a resultados que seriam obtidos em situações reais, com maior variedade de aparências de pessoas e presença de objetos variados no recinto. Porém, o uso daquelas imagens viabilizou a realização do trabalho, eliminando a dependência em relação a pessoas reais a serem usadas em experimentos ou a se encontrar um banco de imagens que tivesse muitas imagens de um mesmo recinto, sob um mesmo ângulo, rotuladas com o número exato de pessoas em cada exemplo, incluindo as pessoas fora do campo de visão da câmera.

No estudo bibliográfico, os melhores resultados encontrados tiveram resultados de contagem com erros absolutos médios da ordem de entre um e dois. Porém, uma comparação

direta deles com os resultados aqui obtidos não é imediata, devido às diferenças entre os bancos de dados usados tanto quanto às imagens presentes neles, quanto aos seus números de imagens. Outra importante diferença reside nos objetivos dos trabalhos: enquanto aqueles almejaram contar as pessoas visíveis nas imagens, este buscou estimar o número de pessoas presentes num recinto, incluindo aquelas que não estavam no campo de visão da câmera.

Para trabalhos futuros, recomenda-se que seja levada em consideração a melhoria significativa na densidade e iluminação do ambiente. O problema de sobreposição de indivíduos na estimativa de multidões pode ser reduzido criando um modelo YOLO que detecte apenas cabeças em capturas de imagens, em vez de todo o corpo, como foi feito neste estudo. Além disso, a validação do sistema em ambientes reais, com a captura de imagens em tempo real, é crucial para avaliar sua eficácia em situações dinâmicas. A integração de métodos de correção de distorção e o aprimoramento da resolução das imagens podem contribuir para reduzir possíveis distorções nas estimativas.

Este trabalho possibilitou testar diversas técnicas para se obter resultados satisfatórios. Um estimador de número de pessoas sempre será necessário para a população, seja para saber se um determinado transporte público está ou não lotado, como o controle de pessoas em ambientes hospitalares, pois, diante da pandemia de 2020, a população está cada vez mais consciente do quão necessário é estar em distanciamento social e controle do número de pessoas em eventos.

Agradecimentos

Os autores agradecem aos professores Miguel Angelo de Abreu de Sousa e Sara Dereste dos Santos, pelas sugestões dadas na escrita do artigo.

Referências

Amazon. What is Neural Network? 2023. Disponível em: https://aws.amazon.com/what-is/neural-network/#:~:text=A%20neural%20network%20is%20a,that%20resembles%20the%20human%20brain. Acesso em 29 out 2023.

Blender Foundation. Artistic Freedom Starts with Blender. Disponível em https://www.blender.org/. Acesso em 11/11/2023.

Conselho Nacional De Justiça (CNJ). A superlotação carcerária é um fenômeno histórico, persistente e caro no Brasil. 2022. Disponível em: https://www.cnj.jus.br/wp-content/uploads/2022/03/folder-central-regulacao-vagas.pdf. Acesso em: 19 ago. 2023.

Eaton, J. W. GNU Octave. Disponível em https://octave.org/. Acesso em 14/11/2023.

Folha De São Paulo. Metrô de SP é mais lotado que o de Tóquio. São Paulo, 21 abr. 2008. Disponível em: https://www1.folha.uol.com.br/fsp/cotidiano/ff2104200810.htm. Acesso em: 19 ago. 2023.

Freund, Rudolf J.; Wilson, William J.; Sa, Ping. **Regression analysis**. Elsevier, 2006.

G1. Alemanha investiga causas do tumulto que matou 19 na Love Parade. 2010. Disponível em: https://g1.globo.com/mundo/noticia/2010/07/alemanha-investiga-causas-do-tumulto-que-matou-19-na-love-parade.html. Acesso em: 19 ago. 2023.

G1. Dois anos depois: veja 24 erros que contribuíram para a tragédia na Kiss. Rio Grande do Sul, 27 jan. 2015. Disponível em: https://g1.globo.com/rs/rio-grande-do-sul/noticia/2015/01/dois-anos-depois-veja-24-erros-que-contribuiram-para-tragedia-na-kiss.html. Acesso em: 19 ago. 2023.

Galimberti, Luiz. Estudo Comparativo de algoritmos de Biometria Facial disponibilizados pela biblioteca OpenCV para controle de acesso. Universidade do Vale do Taquari. Lageado. Dezembro de 2018.

Gonzalez, R.; Woods, R. E. Digital Image Processing. Pearson; 4a edição. ISBN 9780133356724. 2017.

Gouiaa, Rafik; Akhloufi, Moulay A.; Shahbazi, Mozhdeh. Advances in convolution neural networks based crowd counting and density estimation. **Big Data and Cognitive Computing**, v. 5, n. 4, p. 50, 2021.

Haykin, S. Neural networks: a comprehensive foundation. Prentice-Hall, Inc., 2007.

IBGE. Cidades. São Paulo: Panorama. 2022. Disponível em: https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama. Acesso em: 19 ago. 2023.

Ilyas, Naveed; Shahzad, Ahsan; Kim, Kiseon. Convolutional-neural network-based image crowd counting: Review, categorization, analysis, and performance evaluation. **Sensors**, MDPI. v. 20, n. 1, p. 43, 2019.

Keras. Keras: Simple, Flexible, Powerful. Disponível em https://keras.io/. Acesso em 15/11/2023.

Li, Bo et al. Approaches on crowd counting and density estimation: a review. **Pattern Analysis and Applications**, v. 24, p. 853-874, 2021.

Moraes, Pedro Henrique. Contagem do Fluxo de Pessoas Utilizando Aprendizado Profundo. Fundação Universidade Federal de Mato Grosso do Sul. Campo Grande. Agosto de 2021.

OpenCV Team. OpenCV. 2023. Disponível em https://opencv.org/. Acesso em 13/11/2023.

Pinto, Ellen. Shiguemor, Elcio. Vijaykumar, Nandamudi. Detecção Automática de pessoas com uso de Redes Neurais Convolucionais para aplicações de cálculo de trajetória de drones. INPE - Florianópolis - SC. Abril de 2023.

Sindagi, V. A.; Patel, V. M. A survey of recent advances in CNN-based single image crowd counting and density estimation. Pattern Recognition Letters. 107. 3-16. 2018.

Taunk, Kashvi et al. A brief review of nearest neighbor algorithm for learning and classification. In: **2019 international conference on intelligent computing and control systems (ICCS)**. IEEE, 2019. p. 1255-1260.

Tewari, Ayush et al. State of the art on neural rendering. In: **Computer Graphics Forum**. 2020. p. 701-727.

The Linux Foundation. Ultralytics YOLOv5 for object detection, instance segmentation and image classification. Disponível em https://pytorch.org/hub/ultralytics_yolov5/. Acesso em 02/11/2023.

Vargas, Ana; Paes, Aline; Vasconcelos, Cristina. Um Estudo sobre Redes Neurais Convolucionais e sua Aplicação em Detecção de Pedestres. Instituto de Computação Universidade Federal Fluminense. Niterói. Outubro de 2016.

Veja São Paulo. Entenda como são feitos os cálculos de público da PM e do Datafolha. 2016. Disponível em: https://vejasp.abril.com.br/cidades/calculos-publico-policia-militar-datafolha Acesso em 24/08/2023.