Traffic Flow Classification using a Video Descriptor and a Convolutional Neural Network

Matheus Vieira Lessa Ribeiro
Departamento de Engenharia Elétrica
Universidade Federal do
Espírito Santo
Vitória - ES, Brazil
matheusvlribeiro@gmail.com

Jorge Leonid Aching Samatelo
Departamento de Engenharia Elétrica
Universidade Federal do
Espírito Santo
Vitória - ES, Brazil
jorge.samatelo@ufes.br

***Resumo*—Traffic congestion is a significant problem in urban cities and affects economic, health, and social questions. Although many works have been published in the last years to traffic applications based on video data, different techniques of computer vision can be explored in this area. In this work, we proposed a method for traffic flow classification using StarRGB and Convolutional Neural Networks (CNN). The StarRGB describes a global representation of the traffic video into a colored image based on motion elements in the scene. Then, the generated image passed as input to a pre-trained CNN to extract the features and classify the traffic video activity in three classes: LIGHT, MEDIUM, and HEAVY. In our experiments using a traffic video database, the proposed method reached an accuracy of 96.47%. Also, the results suggest that StarRGB is a good descriptor for traffic video applications.**

***Index Terms*—Computer Vision, StarRGB, CNN, Traffic Flow Classification**

I. Introdução

O fluxo de carros nas cidades brasileiras tem sido um dos principais problemas enfrentados pelas autoridades municipais ao longo dos últimos anos. Entre os fatores que levaram esta situação, é possível citar o aumento da frota de veículos, a falta de planejamento urbano frente ao crescimento da densidade demográfica nestas cidades, o comportamento dos motoristas ao cometerem infrações, e a deficiência de um transporte público de qualidade [1].

O congestionamento devido ao elevado fluxo de veículos pode proporcionar perdas econômicas na ordem de bilhões de dólares anualmente [2]. Além disso, está relacionado com os sintomas de estresse , sendo uma das causas de doenças psicológicas e ações violentas por parte dos motoristas ultimamente [3].

Neste contexto, a detecção automática da condição do trânsito nas vias urbanas é uma tarefa de importância, já que permite alertar aos motoristas dos pontos de alta incidência de veículos e consequentemente reduzindo o tempo de trajetória, bem como o custo do combustível utilizado para o deslocamento.

Um sistema capaz de identificar em tempo real as vias com alta taxa de veículos também oferece outras vantagens tanto para o ambiente quanto para a sociedade. A redução da poluição causada pela emissão de gás carbônico pelos veículos, contribui no combate aos problemas referentes ao aquecimento global. Além disso, aumenta-se a eficiência dos sistemas de transporte e há uma maior rapidez para atendimento de chamadas de emergência [4] [5] [6].

Visando criar aplicações para os sistemas inteligentes de transporte (*Intelligent Transportation Systems* -ITS), cientistas e pesquisadores estão preferindo utilizar imagens e vídeos gerados pelas câmeras de monitoramento. Entre os principais motivos estão a infraestrutura já disponibilizada em virtude do número de câmeras instaladas nas vias públicas, a área de alcance destas câmeras, e o aumento de técnicas de visão computacional e processamento de imagens [7] [8] [2].

À vista disso, vários objetivos podem ser atingidos através da análise de imagens e vídeos de trânsito, como a detecção de infração e acidentes, o gerenciamento de semáforos em cruzamentos, a identificação e contagem de veículos e a classificação automática de fluxo de carros [4] [9].

A classificação do fluxo de carros consiste em rotular a condição do trânsito de veículos na via, considerando algumas variáveis como o número de carros em um curto espaço de tempo e a velocidade média deles [10]. Em geral, na literatura há trabalhos classificando o trânsito em duas, três ou quatro classes distintas [11] [12] [13].

Entretanto, este cenário possui uma série de dificuldades e desafios como a variedade de resolução e escala para cada câmera instalada, o alto índice de oclusão gerado pelos veículos em situações de tráfego intenso, a iluminação variando durante o dia e as condições climáticas como a chuva, que atrapalham a detecção de carros [14].

Desta forma, este trabalho propõe um novo modelo para classificação de trânsito através da análise de vídeos, utilizando técnicas de detecção de movimentos, através do algoritmo StarRGB [15]. Além disso, a extração e classificação das características da resposta do StarRGB é realizada por uma *Convolutional Neural Network* (CNN).

Portanto, esta pesquisa possui duas contribuições: (*i*) validar o algoritmo StarRGB como um bom descritor para a análise de vídeos de trânsito; (*ii*) utilizar uma CNN com seus pesos já treinados e aplicá-la à imagem gerada pelo StarRGB para classificar o fluxo de trânsito do vídeo em três classes.

O presente artigo está organizado da seguinte forma, no

capítulo I é feita uma introdução a respeito da classificação de trânsito; no capítulo II são descritas as principais técnicas utilizadas na literatura para solucionar este problema; no capítulo III nossa proposta é detalhada especificando os métodos utilizados; no capítulo IV são mostrados os resultados através dos experimentos realizados em uma base de dados e no capítulo V são expostas as conclusões e trabalhos futuros.

II. Trabalhos Relacionados

Durante os últimos anos, diferentes técnicas de visão computacional tem sido apresentadas com o objetivo de classificar ou estimar o fluxo de trânsito em uma via. As primeiras abordagens exploraram algoritmos computacionais para segmentar, detectar e rastrear os carros de acordo com os movimentos realizados no vídeo. Entre as principais técnicas o *Background Subtraction*(BS) recebeu grande atenção [16] [8] [17].

Todavia, a detecção e o rastreamento de múltiplos objetos simultaneamente enfrenta vários desafios, principalmente em situações de congestionamento onde há um grande número de carros em baixa velocidade. Além disso, o rendimento depende da resolução e qualidades das câmeras de videomonitoramento [18]. De modo a solucionar este problema, diferentes projetos tem sido propostos utilizando abordagens holísticas. Neste caso, a imagem é tratada como um único objeto com o propósito de extrair informações de natureza global a respeito da cena, sem a necessidade de implementar outras técnicas como segmentação e rastreamento [19].

Neste contexto, alguns trabalhos foram desenvolvidos extraindo características referentes à textura nos vídeos de trânsito [12] [20].

Chan et al. [12] propuseram um modelo para determinar a condição atual do trânsito através de um processo estocástico auto-regressivo, este algoritmo codifica as componentes espaciais e temporais de textura na cena em duas distribuições de probabilidade. Enquanto que Derpanis e Wildes [20] associaram cada vídeo com um histograma de orientações de espaço-tempo através de um conjunto de filtros derivativos da função gaussiana. Ambos trabalhos classificaram a condição do tráfego de veículos na via em três classes diferentes.

Em [11], o algoritmo GLCM (*Gray-Level Co-Ocurrence Matrix*) foi aplicado para representar as características horizontais, verticais e diagonais de textura na imagem na detecção de congestionamento da via.

Com o recente avanço do poder computacional para o processamento e armazenamento de dados, o uso das CNNs melhorou o rendimento em diversos campos de aplicação [21]. Por conseguinte, o emprego dessas técnicas nas abordagens holísticas envolvendo classificação e previsão do trânsito recebeu grande atenção nos últimos anos [13] [7] [18] [22].

Luo et al. [19] investigaram um método que não utiliza informações de movimento na cena. Para adquirir uma representação global do tráfego, os autores testaram quatro descritores visuais e dois modelos de CNN treinados no banco de dados ImageNet [23]. Em uma versão estendida para processar videos com baixa taxa de amostragem, Luo et al. [18] testaram um conjunto de CNNs e obtiveram melhores resultados. Os autores extraíram informações do vídeo a partir de diferentes camadas das redes utilizadas e usaram o classificador *Support Vector Regression*-SVR para classificar o fluxo de trânsito em três classes possíveis.

Para classificar imagens de trânsito em quatro classes distintas, Pamula [13] apresentou um método utilizando um conjunto de filtros ótimos para a extração de textura e uma CNN para a classificação. Em [7] é proposto um método inspirado no comportamento humano para contar o número de veículos em uma cena sem a necessidade de detectar cada veículo individualmente. A proposta utiliza apenas uma CNN para capturar a informação global do vídeo.

De modo geral, os melhores resultados obtidos na área de classificação de fluxo de trânsito e detecção de congestionamento utilizam as CNNs em sua abordagem holística. Neste contexto, trabalho presente propõe o uso destas redes para classificar vídeos de monitoramento de trânsito em três classes. Inicialmente,um algoritmo para representação e compactação de vídeos em imagens coloridas é utilizado de modo a gerar uma imagem de entrada para a CNN.

III. Proposta

A Figura 1 apresenta a proposta deste trabalho. Em suma, a abordagem é dividida em duas etapas: pŕe-processamento e classificação. Inicialmente, na etapa de pré-processamento o vídeo é compactado e transformado em uma imagem RGB (*Red, Green, Blue*) através do algoritmo StarRGB [15]. Então, a imagem gerada é aplicada a uma CNN que extrai suas principais características e classifica a condição do trânsito presente no vídeo em uma de três classes distintas: LIGHT, MEDIUM e HEAVY. Em nossa proposta, a CNN utilizada foi a VGG16 [24]. Cada uma destas etapas será explicada em detalhe nos próximos parágrafos.

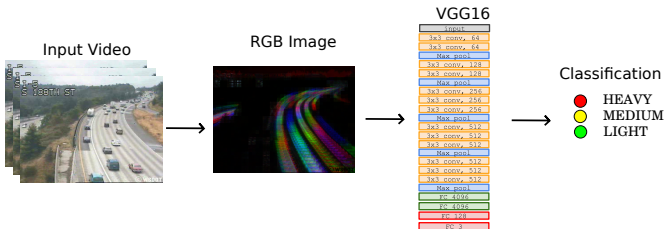

Figura 1: Framework da nossa proposta para classificação do fluxo de trânsito.

A. Pré-Processamento

O algoritmo StarRGB foi construído com o objetivo de descrever vídeos contendo gestos dinâmicos e representá-los em uma única imagem RGB [15]. A principal motivação para a elaboração da proposta pelos autores foi contribuir com a representação original proposta em [25], adicionando informações de cor e componentes temporais ao modelo de reconhecimento de gestos dinâmicos.

Seja uma matriz RGB $\mathbf{I}_k$ representando o k^{th} frame RGB de um vídeo e seja a matriz $\mathbf{M}$ a representação final de um vídeo pelo algoritmo Star original proposto em [25]. Para avaliar

as mudanças de intensidade, cromaticidade e saturação entre *frames* consecutivos, Santos et al. [15] modificaram o cálculo da matriz $\mathbf{M}$ utilizando a métrica proposta em [26], baseada na similaridade dos cossenos, representadas pela Equações (1) e (2).

$$\lambda = 1 - \cos(\theta) = 1 - \frac{\mathbf{I}_{k-1}(i,j)^T \mathbf{I}_k(i,j)}{||\mathbf{I}_{k-1}(i,j)||_2 ||\mathbf{I}_k(i,j)||_2}, \quad (1)$$

$$\mathbf{D}_k(i,j) = (1 - \frac{\lambda}{2}).|||\mathbf{I}_{k-1}(i,j)||_2 - ||\mathbf{I}_k(i,j)||_2|. \quad (2)$$

em que (i,j) representa as posições na matriz $\mathbf{I}$ e θ é o ângulo entre $\mathbf{I}_{k-1}(i,j)$ e $\mathbf{I}_k(i,j)$. Desta forma, para N *frames* consecutivos, a matriz $\mathbf{M}$ é calculada pela seguinte equação:

$$\mathbf{M}(i,j) = \sum_{k=2}^{N} \mathbf{D}_k(i,j). \quad (3)$$

Em suma, o algoritmo StarRGB possui duas etapas para a construção da representação final de um vídeo em uma imagem RGB:

- Cada vídeo com N *frames* é dividido em três sub-vídeos, contendo cada sub-vídeo o mesmo número de *frames*. Caso o número total de *frames* não seja divisível por três, o sub-vídeo central irá conter $N - 2\lfloor N/3 \rfloor$ *frames*.
- Para cada sub-video, a matriz $\mathbf{M}$ é calculada via a Equação 3. Desta forma, a matriz $\mathbf{M}$ obtida de cada sub-video compõe um canal na imagem RGB construída. Sendo assim, o canal R (*Red*) representa a matriz $\mathbf{M}$ calculada a partir do primeiro sub-video, o canal G (*Green*) representa a matriz $\mathbf{M}$ do sub-video central e o canal B (*Blue*) representa a matriz $\mathbf{M}$ do terceiro sub-video.

O referido algoritmo aumentou o poder descritivo da representação Star de acordo com os resultados obtidos em [15]. Outrossim, a informação de movimento nos vídeos de trânsito é relevante para a descrição da condição do fluxo na via [27] [8]. Portanto, de modo a descrever os movimentos dos carros, o algoritmo descrito será testado em vídeos de trânsito para gerar uma nova representação por meio de uma imagem RGB. A Figura 2 ilustra a representação StarRGB adquirida a partir de vídeos com diferentes condições de transito na via.

B. Classificação

Após o pré-processamento do video de trânsito para a construção da imagem RGB correspondente, o próximo desafio é classificar esta representação em três classes distintas: LIGHT, MEDIUM e HEAVY. Em nossa abordagem, uma CNN foi utilizada para esta etapa, tendo como entrada a imagem gerada pelo StarRGB e a saída uma camada com três neurônios e função de de ativação *softmax*.

Basicamente, uma CNN consiste em uma série de camadas empilhadas de quatro tipos específicos: convolucionais, *pooling* e *fully connected* [18]. Há varias arquiteturas diferentes na literatura, cada qual apresentando quantidade diferentes de camadas contendo os quatro tipos citados.

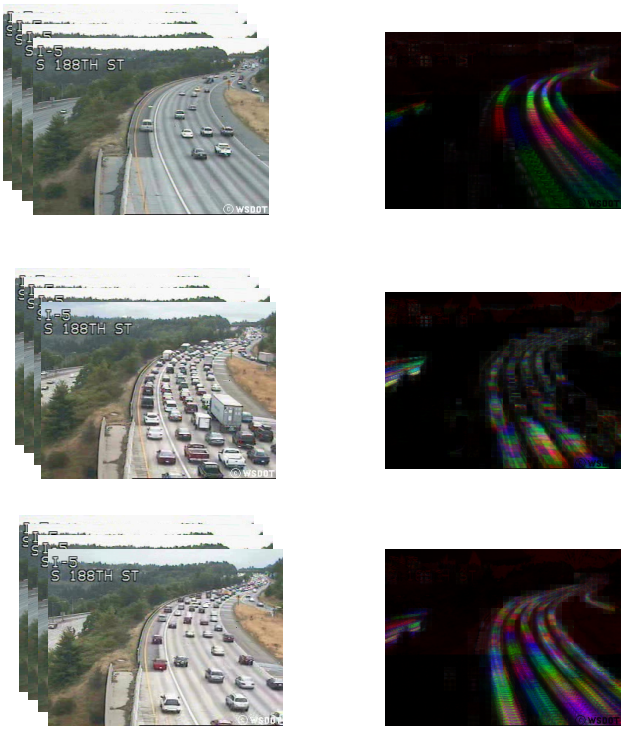

Figura 2: *Frames* de vídeos com diferentes condições de trânsito e suas respectivas representações geradas pelo algoritmo StarRGB.

Em nosso trabalho, a rede escolhida foi a VGG16 proposta por [24]. Em virtude da sua simplicidade e de vários trabalhos e aplicações apresentarem bons resultados com a utilização desta arquitetura. É possível encontrar a implementação da VGG16 em aplicações na área de classificação de trânsito, como realizado em [18].

A CNN VGG16 consiste em cinco blocos convolucionais sequenciais e três camadas de *fully connected*. Cada bloco convolucional esta constituído por uma série de camadas convolucionais e *max-pooling*. A CNN VGG16 foi elaborada originalmente para classificar imagens de tamanho 224×224 em 1000 classes [24]. Porém, a rede pode atuar também como um extrator de características de uma imagem com tamanho arbitrário. O processo de extração de características da representação do vídeo de trânsito foi realizado através dos pesos pré-treinados da VGG16 sobre o banco de dados ImageNet, especializado em classificação de imagens. Este banco contém mais de 1.2 milhão de imagens distribuídas em 1000 classes distintas [23].

Finalmente, para classificar a imagem RGB, a camada de saída da rede VGG16 com 1000 neurônios foi substituída por uma sequência de duas camadas *fully connected* a primeira com 128 neurônios e função de ativação *relu*, e a segunda com 3 neurônios e função de ativação *softmax*. A rede foi treinada sobre o banco de dados UCSD explicado em detalhe na seção IV-A

Figura 3: Classes de fluxo de trânsito do banco de dados UCSD.

IV. Resultados e Discussão

Nesta seção é descrito o banco de dados utilizado para validação da proposta apresentada na seção anterior, bem como os resultados dos experimentos realizados.

A. Banco de dados

O banco de dados UCSD[1] contém 254 vídeos de monitoramento de uma rodovia da cidade de Seattle, gravados pelo Departamento de Transporte do Estado de Washington, nos Estados Unidos[2]. Todo o período de gravação possui aproximadamente 20 minutos em uma taxa de 10 *frames* por segundo. Cada vídeo possui 4 a 5 segundos de duração e uma resolução de 320×240 [12].

O banco de dados inclui três classes de fluxo de trânsito de acordo com o *ground-truth* rotulado manualmente: LIGHT, MEDIUM e HEAVY. A classe LIGHT, com 165 amostras, corresponde a um fluxo de trânsito com baixo número de veículos percorrendo em velocidades próximas ao limite estabelecido da via. De modo contrário, a classe HEAVY, com 44 amostras, representa a situação de trânsito intenso e congestionamento. Nos vídeos pertencentes a esta classe há uma quantidade maior de carros se deslocando com uma velocidade baixa na via. Por fim, a classe MEDIUM, com 45 amostras, corresponde a um meio termo entre as duas classes apresentadas. A Figura 3 ilustra um *frame* de um vídeo deste banco de dados, correspondente a cada classe.

B. Experimentos

A metodologia para treinamento e teste aqui é a mesma usada em [12] e em outros trabalhos na literatura que utilizam este conjunto de vídeos. O banco de dados é dividido em 75% para treino e 25% para teste. Quatro ensaios são realizados e o próprio UCSD disponibiliza a sequência de vídeos para o conjunto de treino e teste em cada ensaio. A acurácia Acc de cada ensaio é definida como $Acc = \frac{CC}{CC+CE}$, em que *CC* (Classe Correta) é o número de vídeos classificados corretamente e *CE* (Classe Errada) é o número de vídeos classificados de maneira incorreta. A acurácia final é a média aritmética da acurácia dos quatro ensaios realizados.

A máquina utilizada para execução dos experimentos contém a seguinte configuração: (*i*) Sistema Operacional Linux Ubuntu Server, versão 18.04; (*ii*) Processador Intel(R) Core(TM) i7-2600, 3.40GHz com 4 núcleos físicos; (*iii*) 16GB de memória RAM; (*iv*) 500GB de armazenamento (disco rígido); (*v*) GPU (*Global Processing Unit*) Nvidia GTX 1080Ti, com $11GB$ de memória RAM.

Inicialmente as imagens geradas pelo StarRGB foram normalizadas dividindo a intensidade dos pixeis por 255. Utilizando o processo de *Transfer Learning* para a VGG16, congelou-se os pesos pré-treinados pelo ImageNet, e as duas últimas camadas de *fully connected* adicionadas foram retreinadas.

Para este treinamento foram adotados os seguintes hyperparâmetros: semente com valor 31; *batch size* igual a 32; máximo de 100 épocas; taxa de aprendizagem de 0.001 e *momentum* de 0.9; o algoritmo de otimização escolhido foi o gradiente descendente estocástico padrão.

A Tabela I representa a acurácia final obtida com os experimentos, além da acurácia de cada um dos quatro ensaios (T) correspondentes. Com os testes realizados, a performance adquirida foi de 96.47%.

A matriz de confusão acumulativa para os resultados apresentados na Tabela I é descrita na Tabela II. Ao todo, 9 vídeos de 254 foram classificados incorretamente. A classe LIGHT obteve a melhor performance em comparação às outras

[1]disponível em http://www.svcl.ucsd.edu/projects/traffic/
[2]https://www.wsdot.wa.gov/

Figura 4: Amostras pertencentes à classe HEAVY incorretamente classificados como MEDIUM

classes, com apenas uma amostra classificada de maneira errada. Observe que o maior percentual de erro está na classificação dos vídeos da classe HEAVY como MEDIUM. De fato, a variância entre estas duas classes é pequena, como pôde ser observado na Figura 3, isto prejudicou a classificação. A Figura 4 mostra um *frame* de cada um dos cinco vídeos incorretamente classificados desta forma.

Os resultados apresentados são comparados com outras propostas presentes na literatura para este mesmo banco de dados, na Tabela III. A melhor acurácia pertence ao trabalho recente de Luo et al. [18], em que há uma série de testes extraindo as informações de várias camadas de dois modelos de CNN para segmentar as regiões na imagem de trânsito.

Embora nossa proposta não obteve a melhor performance, os resultados sugerem que o estudo referente à implementação de outras CNNs ou de outras camadas para a saída da VGG16 como extrator de características, podem aumentar a acurácia. O algoritmo StarRGB mostrou ter um bom compromisso em descrever e representar vídeos de trânsito, visto que utilizando apenas a VGG16 sobre as imagens geradas, os resultados superaram vários trabalhos na literatura.

V. CONCLUSÃO

Neste trabalho, foi proposta uma nova abordagem para representar e classificar vídeos de trânsito nas classes LIGHT, MEDIUM e HEAVY de acordo com o fluxo de veículos na via. Para isto, a proposta é dividida em duas partes, a primeira utiliza um algoritmo para representação de vídeos em imagens coloridas. Já a segunda parte consiste na extração e classificação das características desta imagem por uma CNN.

Para compactar e transformar o vídeo em uma imagem RGB, o algoritmo StarRGB foi implementado. O método procura os movimentos na cena considerando a diferença entre pixeis subsequentes. Não obstante, o StarRGB acrescenta informação de cor e informação temporal à imagem representada, aumentando seu poder descritivo.

A extração de características das imagens geradas sua pelo StarRGB e sua classificação foi feita por uma rede CNN, a VGG16. Desta forma, os pesos treinados pelo banco de dados ImageNet foram mantidos e a ultima camada substituída para retreinar os seus pesos, usando *Transfer Learning*.

Os resultados obtidos após experimentos sobre o banco de dados UCSD superou vários trabalhos já consolidados na literatura que utilizam este mesmo banco de dados. Os testes sugerem que o StarRGB pode ser um bom descritor para vídeos de trânsito, em que o movimento dos carros deve ser considerado.

Para projetos futuros será estudado a implementação de outras CNNs ou a utilização de outras camadas da VGG16 para extração de características.

Tabela I: Resultados obtidos pelos experimentos com o banco de dados UCSD.

Total	**T1**	**T2**	**T3**	**T4**
96.47%	96.83%	95.31%	95.31%	98.41%

Tabela II: Matriz de confusão para os experimentos.

		Previsto		
		LIGHT	MEDIUM	HEAVY
Correto	LIGHT	164	1	0
	MEDIUM	2	42	1
	HEAVY	0	5	39

Tabela III: Resultados de diferentes propostas utilizando o banco de dados UCSD.

Trabalho	*Acc* (%)
Chan and Vasconcellos [12]	94.50%
Andrews,Sobral et al. [16]	94.50%
Riaz and Khan [27]	95.28%
Derpanis and Wildes [20]	95.30%
Asmaa et al. [8]	96.37%
Luo et al. [19]	96.9%
Luo et al. [18]	**97.64%**
Nossa proposta	96.47%

AGRADECIMENTOS

Este projeto é financiado pelas agências brasileiras MCTI/MC/ CGI e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), bolsas 2015/24423-3 e 2018/05150-4. Os autores também reconhecem o apoio da NVIDIA *Corporation* através da doação da GPU usada nesta pesquisa.

REFERÊNCIAS

[1] S. Rahane and U. Saharkar, "Traffic congestion - causes and solutions: A study of talegaon dabhade city," *J. Inf. Knowl. Res. Civil Eng*, vol. 3, pp. 160–163, 01 2014.

[2] P. Chakraborty, Y. O. Adu-Gyamfi, S. Poddar, V. Ahsani, A. Sharma, and S. Sarkar, "Traffic congestion detection from camera images using deep convolution neural networks," *Transportation Research Record.*

[3] L.-P. Beland and D. A. Brent, "Traffic and crime," *Journal of Public Economics*, vol. 160, pp. 96–116, 2018.

[4] A. M. de Souza, C. A. Brennand, R. S. Yokoyama, E. A. Donato, E. R. Madeira, and L. A. Villas, "Traffic management systems: A classification, review, challenges, and future perspectives," *International Journal of Distributed Sensor Networks*, vol. 13, no. 4, p. 1550147716683612, 2017.

[5] P. Borkar and L. G. Malik, "Review on vehicular speed, density estimation and classification using acoustic signal." *International Journal for Traffic & Transport Engineering*, vol. 3, no. 3, 2013.

[6] N. Lefebvre, X. Chen, P. Beauseroy, and M. Zhu, "Traffic flow estimation using acoustic signal," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 164 – 171, 2017.

[7] J. Chung and K. Sohn, "Image-based learning to measure traffic density using a deep convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1670–1675, May 2018.

[8] O. Asmaa, K. Mokhtar, and O. Abdelaziz, "Road traffic density estimation using microscopic and macroscopic parameters," *Image and Vision Computing*, vol. 31, no. 11, pp. 887–894, 2013.

[9] N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *Trans. Intell. Transport. Sys.*, vol. 12, no. 3, pp. 920–939, Sep. 2011.

[10] R. Loce, R. Bala, and M. Trivedi, *Computer Vision and Imaging in Intelligent Transportation Systems*, 04 2017.

[11] L. Wei and D. Hong-Ying, "Real-time road congestion detection based on image texture analysis," *Procedia engineering*, vol. 137, pp. 196–201, 2016.

[12] A. B. Chan and N. Vasconcelos, "Classification and retrieval of traffic video using auto-regressive stochastic processes," in *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*. IEEE, 2005, pp. 771–776.

[13] T. Pamula, "Road traffic conditions classification based on multilevel filtering of image content using convolutional neural networks," *IEEE Intelligent Transportation Systems Magazine*, vol. 10, no. 3, pp. 11–21, 2018.

[14] M. Manana, C. Tu, and P. A. Owolawi, "A survey on vehicle detection based on convolution neural networks," in *Computer and Communications (ICCC), 2017 3rd IEEE International Conference on*. IEEE, 2017, pp. 1751–1755.

[15] C. C. dos Santos, J. L. A. Samatelo, and R. F. Vassallo, "Dynamic gesture recognition by using cnns and star rgb: a temporal information condensation," *CoRR*, vol. abs/1904.08505, 2019.

[16] L. O. Andrews Sobral, L. Schnitman, and F. De Souza, "Highway traffic congestion classification using holistic properties," in *10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications*, 2013.

[17] S. Hu, J. Wu, and L. Xu, "Real-time traffic congestion detection based on video analysis," *Journal of Information and Computational Science*, vol. 9, no. 10, pp. 2907–2914, 2012, cited By 14.

[18] Z. Luo, P.-M. Jodoin, S.-Z. Su, S.-Z. Li, and H. Larochelle, "Traffic analytics with low-frame-rate videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 4, pp. 878–891, 2018.

[19] Z. Luo, P.-M. Jodoin, S.-Z. Li, and S.-Z. Su, "Traffic analysis without motion features," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3290–3294.

[20] K. G. Derpanis and R. P. Wildes, "Classification of traffic video based on a spatiotemporal orientation analysis," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*. IEEE, 2011, pp. 606–613.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[22] D. Jo, B. Yu, H. Jeon, and K. Sohn, "Image-to-image learning to predict traffic speeds by considering area-wide spatio-temporal dependencies," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1188–1197, 2019.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] P. Barros, G. I. Parisi, D. Jirak, and S. Wermter, "Real-time gesture recognition using a humanoid robot with a deep neural architecture," in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 646–651.

[26] J. L. A. Samatelo and E. O. T. Salles, "A new change detection algorithm for visual surveillance system," *IEEE Latin America Transactions*, vol. 10, no. 1, pp. 1221–1226, 2012.

[27] A. Riaz and S. A. Khan, "Traffic congestion classification using motion vector statistical features," in *Sixth International Conference on Machine Vision (ICMV 2013)*, vol. 9067. International Society for Optics and Photonics, 2013, p. 90671A.