Uma Abordagem baseada em Redes Neurais, *Multiple Instance Learning* e PCA para Detecção de Anomalias em Videovigilância

Silas S. L. Pereira[1,2], **J. E. Bessa Maia**[1]

[1]Universidade Estadual do Ceará – CCT-PPGCC-UECE
Av. Paranjana, 1700 – 60740-903 – Fortaleza, CE – Brasil

[2]Instituto Federal de Educação, Ciência e Tecnologia do Ceará – Campus Aracati
Aracati – CE – Brasil

silas.santiago@ifce.edu.br, jose.maia@uece.br

Resumo. *Multiple Instance Learning (MIL) tem se tornado uma solução atrativa na literatura de videovigilância por permitir lidar com bases fracamente rotuladas. Este trabalho propõe e avalia uma abordagem para detecção de anomalias em vídeo baseada em classificação binária com redes neurais Multilayer Perceptron (MLP) e paradigma MIL. Os experimentos foram conduzidos a partir de um conjunto de atributos I3D (Inflated 3D) referentes ao dataset de benchmark ShanghaiTech. Explora-se ainda o efeito da compacticidade dos dados e representação de informação essencial com a técnica de extração de atributos Principal Component Analysis (PCA). Os resultados alcançados foram competitivos quando comparados com o estado da arte.*

palavras-chave: detecção de anomalia, videovigilância, Multiple Instance Learning, atributos I3D, Multilayer Perceptron.

Abstract. *Multiple Instance Learning (MIL) has become an attractive solution in video surveillance literature once it allows working with weakly supervised bases. This work proposes and evaluates a video anomaly detection approach based on binary classification with Multilayer Perceptron (MLP) neural networks and MIL paradigm. The experiments were performed from a set of I3D (Inflated 3D) features which corresponds to the benchmark dataset ShanghaiTech. We also explore the effect of compactness and essential data representation with the feature extraction technique Principal Component Analysis (PCA). The achieved results were competitive when compared with state of art.*

keywwords: anomaly detection, video surveillance, Multiple Instance Learning, I3D features, Multilayer Perceptron.

1. Introdução

Detecção de anomalias é uma tarefa crucial em diferentes domínios de aplicação, como detecção de intrusão, fraudes e videovigilância [Li et al. 2021, Pereira and Maia 2021]. Devido à subjetividade na definição de anomalias, estas podem ser de difícil definição, dependem da localização e situação e variam amplamente quanto ao conteúdo e duração [Wan et al. 2020]. O entendimento do contexto é fundamental, visto que uma atividade

anômala em um dado contexto pode ser considerada normal em outro [Rao et al. 2017]. Um evento anômalo é considerado uma situação rara, de modo que possui baixa probabilidade de ocorrência entre a maioria dos eventos normais. Esta característica dificulta a obtenção de exemplos anômalos, visto que a maior parte dos dados reflete padrões de comportamento normal [Rao et al. 2017]. Paradigmas de classificação unária e binária são tipicamente utilizados em abordagens de detecção de anomalias em vídeo apresentadas na literatura. Aprendizagem fracamente supervisionada mitiga a dificuldade inerente à rotulação de instâncias, uma etapa laboriosa durante a construção dos modelos preditivos. Tal paradigma é geralmente formulado como um problema de aprendizagem de múltiplas instâncias (*Multiple Instance Learning* – MIL) [Wan et al. 2020]. MIL é uma abordagem útil em problemas nos quais o conhecimento sobre os rótulos e exemplos de treinamento é incompleto [Ali and Shah 2008]. Em MIL, *bags* contendo múltiplas instâncias, em vez de instâncias individuais, são rotuladas nas classes normal ou anomalia e são utilizadas para o treinamento de uma técnica apropriada de aprendizagem de máquina. A abordagem MIL tem apresentado eficiência e ganhos de desempenho em tarefas de detecção e reconhecimento em trabalhos recentes [Yun et al. 2012].

A extração de características relevantes dos dados é um aspecto essencial para o bom desempenho de métodos de detecção de anomalias[Ribeiro et al. 2018]. Abordagens de aprendizagem profunda (*deep learning*) têm alcançado resultados notáveis em diferentes domínios de aplicação, como classificação de imagens, detecção de objetos, processamento de voz e detecção de anomalias [Pawar and Attar 2019]. Neste trabalho, o conceito de anomalia é compreendido como um evento espaço-temporal cujos padrões se distinguem daqueles comumente encontrados no ambiente observado. Um exemplo de atividade anômala a ser detectada é a ocorrência de ciclismo ilegal em uma via para trânsito exclusivo de pedestres.

Neste trabalho, a tarefa de detecção de anomalias em videovigilância é modelada como um problema de classificação binária usando redes neurais artificiais *Multilayer Perceptron* (MLP) [Haykin 2010]. A abordagem MIL é utilizada de modo que apenas os rótulos de vídeo são empregados na fase de treinamento, de forma a mitigar a tarefa de rotulação em nível de *frame* para construção de classificadores. Um conjunto de atributos extraídos com a rede neural profunda *Inflated 3D (I3D)* para o *dataset* de *benchmark ShanghaiTech* é utilizado para modelagem e avaliação. A técnica *Principal Component Analysis* (PCA) [Haykin 2010] é empregada para redução de dimensionalidade e obtenção de ganhos de desempenho dos modelos de detecção. A partir da avaliação e comparação da abordagem proposta com o estado da arte, pode-se verificar resultados promissores.

Este trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados à temática de detecção de anomalias; a Seção 3 descreve a metodologia utilizada para desenvolvimento da pesquisa proposta; na Seção 4, são apresentados e discutidos os resultados alcançados; a Seção 5 apresenta as conclusões e direcionamentos para pesquisa futura.

2. Trabalhos Relacionados

Esta seção é uma breve revisão de alguns trabalhos diretamente relacionados com este, com o objetivo de contextualizar a presente pesquisa na literatura. Artigos de revisão (*surveys*) recentes que expandem a cobertura desta seção podem ser encontrados em

[Suarez and Naval Jr 2020] e [Nayak et al. 2020].

Em [Sultani et al. 2018], os autores propõem uma abordagem para detecção de anomalias em vídeo a partir da utilização de um conjunto fracamente rotulado de vídeos de treinamento. Para a utilização da abordagem fracamente supervisionada, os autores empregam a abordagem MIL. O problema de detecção de anomalias é então explorado como um problema de regressão, no qual um vetor de características é mapeado para um *score* de anomalia. Os vídeos de vigilância normais e anômalos foram segmentados em clipes, de forma que um vídeo (*bag*) contém múltiplos segmentos de vídeo (instâncias de um *bag*). A geração dos *scores* de anomalia para os segmentos do vídeo é realizada a partir do modelo preditivo gerado a partir dos dados. Para treinamento e avaliação da abordagem proposta, empregou-se um *dataset* para detecção de anomalias em vídeo de larga escala composto por diferentes eventos anômalos. A partir dos resultados obtidos, pode-se verificar que a abordagem proposta é capaz de atingir desempenho superior quando comparada com outras abordagens de detecção de anomalias no estado da arte.

Em [Kamoona et al. 2020], os autores propõem uma rede neural profunda *encoding-decoding* para detecção de anomalias em cenários de videovigilância no intuito de permitir a captura de informações temporais e espaciais das instâncias de vídeo. Segundo os autores, a principal contribuição é considerar as relações temporais entre as instâncias de vídeo de forma a tratá-las como dados visuais sequenciais em vez de um conjunto de instâncias independentes. A partir de um *framework* baseado em aprendizagem fracamente supervisionada com abordagem MIL, os autores propõem uma função de custo que penaliza detecções incorretas. Esta função maximiza a distância média entre as predições anômalas e normais. Os autores utilizaram os *datasets* de *benchmark* em videovigilância UCF-Crime e *ShanghaiTech*. Um conjunto de atributos espaço-temporais são extraídos de cada vídeo avaliado por meio de um modelo de rede C3D. O desempenho dos modelos avaliados foi realizado utilizando as informações de *ground truth* em nível de *frame*. Os resultados alcançados são competitivos em relação ao estado da arte nos estudos de simulação.

Em [Wan et al. 2020], os autores abordam a problemática de detecção de anomalias em vídeos como um problema de classificação binária. Os autores exploram a aprendizagem fracamente supervisionada a partir da abordagem MIL para permitir a classificação em nível de segmentos de vídeo, de forma que apenas rótulos de vídeo são considerados durante o treinamento. Para isso, os autores propõem, implementam e avaliam o *framework* AR-NET (*Anomaly Regression Net*) para detecção de anomalias em vídeos. Ademais, duas novas funções de custo (*Dynamic Multiple-Instance Learning Loss* e *Center Loss*) para aprender discriminantes para detecção de anomalias são propostas e avaliadas no trabalho. O modelo pré-treinado *Inception-v1 I3D* (*Inflated 3D*) é usado para extração das características. O mesmo faz uso de informações de aparência (RGB) e de movimento (*Optical Flow*). Utilizando o *dataset ShanghaiTech*, os autores aplicaram o modelo *Inflated 3D* como a extrator de características. Pode-se verificar que o modelo proposto supera em desempenho as abordagens do estado da arte para o *dataset* utilizado.

A partir dos trabalhos revisados, pode-se afirmar que a detecção precisa de anomalias em cenas de vídeo com características espaço-temporais distintas é ainda um grande desafio para modelos do estado da arte. A combinação de *features* profundas seguida da aplicação da técnica PCA e uma rede neural rasa (MLP) em aprendizagem MIL é a

contribuição deste trabalho. Os resultados dos dois últimos artigos [Wan et al. 2020] e [Kamoona et al. 2020] foram utilizados para comparação de desempenho neste trabalho, visto que ambos os trabalhos utilizam o *dataset* experimentado neste estudo.

3. Metodologia

Esta seção descreve os procedimentos envolvidos para preparação de dados, modelagem e avaliação de desempenho aplicada à abordagem proposta para detecção de anomalias em cenas de videovigilância. As etapas aplicadas nesta pesquisa são descritas brevemente nas subseções seguintes.

3.1. Preparação dos Dados

Para realização dos experimentos, o presente trabalho utiliza uma base de dados[1] com padrões de anomalia previamente processados a partir do *dataset ShanghaiTech*, conforme descritos em [Wan et al. 2020]. *ShanghaiTech* foi construído a partir de capturas na universidade *ShanghaiTech* e descreve diferentes condições de iluminação e pontos de vista das câmeras. O conjunto de dados processado e representado como atributos I3D possui 237 exemplos disponíveis, onde 330 são vídeos rotulados como normais e 107 são vídeos rotulados como anômalos. Cada instância deste *dataset* corresponde à uma matriz n x 2048, onde n é uma quantidade variável de segmentos existentes no vídeo e 2048 o número de atributos. Para cada vídeo processado, há um rótulo indicando a categoria correspondente (normal ou anomalia), além dos rótulos de cada *frame* presente no vídeo. Neste estudo, utiliza-se o conjunto de dados I3D para construção e avaliação dos modelos preditivos para detecção de anomalias. A preparação dos dados para modelagem e avaliação de desempenho é feita como descrito a seguir. Inicialmente, carrega-se o o conjunto de $N = 437$ vídeos processados $\mathbf{D} = \{(X_i, y_i, yf_i)\}_{i=1}^{N}$. Há 330 instâncias de vídeo normais (N) e 110 vídeos anômalos (A). Na representação I3D, cada instância X_i é uma matriz n x 2048 (onde n é uma quantidade variável de segmentos em um dado vídeo), $y_i \in \{0, 1\}$ é o rótulo do vídeo e $yf_i = [yf_i^1, \ldots, yf_i^k]$ é uma sequência de k rótulos de *frame* presentes no vídeo, onde $yf_i \in \{0, 1\}$ e k é uma quantidade variável de *frames*. Em seguida, $\mathbf{D}$ é particionado em treinamento e teste. Adotou-se o limiar de 75% dos dados para treinamento posterior via validação cruzada e o restante para teste. A composição dos dados de treinamento e teste é realizada da seguinte forma: Inicialmente, particiona-se $\mathbf{D}$ em $\mathbf{D}_N$ e $\mathbf{D}_A$ os quais representam o conjunto de vídeos das classes Normal (N) e Anomalia (A), respectivamente. Em seguida, as partições $\mathbf{D}_N$ e $\mathbf{D}_A$ são divididas em $\mathbf{D}_N^{treino}$, $\mathbf{D}_N^{teste}$, $\mathbf{D}_A^{treino}$ e $\mathbf{D}_A^{teste}$. Por fim, as partições de treinamento e teste são formadas como $\mathbf{D}_{treino} = \{\mathbf{D}_N^{treino}, \mathbf{D}_A^{treino}\}$ e $\mathbf{D}_{teste} = \{\mathbf{D}_N^{teste}, \mathbf{D}_A^{teste}\}$. Desta forma, ambas as partições são construídas de forma a manter a mesma proporção de vídeos de treinamento e teste dos dados originais processados. O conjunto de treinamento é então formado por 247 vídeos normais e 80 vídeos anômalos. O conjunto de teste possui 83 vídeos normais e 27 vídeos anômalos. Em seguida, as partições de treinamento e teste são dispostas para permitir a construção de um modelo preditivo em nível de segmento. Cada vídeo em $\mathbf{D}_{treino}$ e $\mathbf{D}_{teste}$ é então transformado em n novos segmentos com 2048 atributos. A nova partição de segmentos de treinamento é descrita como $\mathbf{S}_{treino} = \{(S_i, y_i)\}$ onde S_i é um segmento de vídeo e y_i a respectiva classe deste segmento, que é a mesma do vídeo que o contém. De forma similar ao estudo de [Wan et al. 2020], o problema de

[1] https://github.com/wanboyang/Anomaly_AR_Net_ICME_2020

detecção de anomalias é neste estudo tratado como um problema fracamente supervisionado seguido de classificação binária, de forma que apenas os rótulos de vídeo são considerados durante a etapa de construção dos modelos preditivos. No intuito de possibilitar a avaliação de desempenho em nível de *frame*, a partição de segmentos de teste é descrita por $\mathbf{S}_{teste} = \{(S_i, yfs_i)\}$, onde yfs_i é um vetor de 16 rótulos de *frame* obtidos a partir de yf_i. O conjunto de treinamento é formado por 14532 segmentos normais e 2049 segmentos anômalos. O conjunto de teste possui 4983 vídeos normais e 561 vídeos anômalos.

3.2. Modelagem e Avaliação

Neste estudo, emprega-se a técnica PCA (*Principal Component Analysis*) [Haykin 2010] para reduzir o espaço de entrada com 2048 atributos para 624 e 205 componentes principais, às quais explicam aproximadamente 95% e 85% da variância dos dados de treinamento, respectivamente. Conforme [Zhao et al. 2020], técnicas para extração de atributos, para as quais PCA é uma das mais amplamente utilizadas, permitem obter representações mais compactas dos dados que incluam informação essencial para tomada de decisão. PCA é uma abordagem não supervisionada que objetiva rotacionar os eixos do espaço de representação de uma matriz de dados, mantendo a ortogonalidade, de tal forma que, quando os dados são projetados sobre esses novos eixos, a variância explicada dos dados é maximizada em ordem decrescente para alguma sequência das direções percorrida. PCA pode ser calculado a partir de diferentes estruturas tais como as matrizes de covariância ou de correlação dos dados. Este trabalho usou decomposição SVD (*Singular Value Decomposition*) [Saha et al. 2009]. Emprega-se a técnica *K-fold Cross-Validation* ao conjunto de treinamento para estimar o desempenho da rede neural em cada iteração do *cross-validation*. O número de *folds* foi definido como $k = 5$, conforme [Kuhn et al. 2013]. O modelo com menor RMSE (*Root Mean Squared Error*) entre as k iterações da validação cruzada é utilizado nos experimentos para avaliação com o conjunto de teste. Considerou-se arquiteturas de rede *Multilayer Perceptron* (MLP) com uma e duas camadas e com 256 neurônios em cada camada escondida. Os parâmetros adotados neste trabalho foram guiados pela literatura. A escolha de uma rede neural densa para classificação binária, assim como os parâmetros taxa de *dropout*, método de otimização e função de ativação para as camadas intermediárias é baseada em [Wan et al. 2020]. Os demais parâmetros aqui descritos para treinamento da rede foram escolhidos arbitrariamente.

Para uma rede MLP com uma única camada escondida, a saída dos neurônios desta camada é denotada pela expressão ${S_i}^{L2} = D(\max(0, W_{L1} \cdot S_i + b_{L1}))$, onde $D(\cdot)$ corresponde à regularização *Dropout* com taxa de 0.70, que fará o descarte de algumas entradas durante o treinamento da rede no intuito de prevenir o sobreajuste do modelo. Utiliza-se a função RELU (*Rectified Linear Unit*), especificada por $y = \max(0, x)$, para a ativação do neurônio da camada escondida. A saída da rede é descrita por $s_i = 1/(1 + \exp(W_{L2} \cdot S_i + b_{L3}))$, representando por um neurônio com função de ativação sigmoidal. O treinamento da rede é em lotes, de maneira que foi definido o valor arbitrário de 512 para o tamanho do lote durante o treinamento e 1000 épocas para o treinamento dos modelos. O otimizador ADAM (*Adaptative Moment Estimation*) [Kingma and Ba 2014] foi utilizado como uma variante da técnica de descida do gradiente. A função de custo *Binary Cross Entropy* foi empregada por ser popularmente utilizada para construção de modelos para

classificação binária, definida a partir da expressão $H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1-y_i)\cdot\log(1-p(y_i))$, onde $y_i \in \{0,1\}$ e $p(y_i)$ é a probabilidade de ocorrência da classe para o padrão i.

4. Resultados e Discussão

Esta seção apresenta os principais resultados obtidos a partir da avaliação de desempenho da abordagem proposta para detecção de anomalias em videovigilância. Na Tabela 1, apresenta-se os valores de desempenho para 6 modelos MLP com configurações distintas de número de camadas e espaço de entrada. Considera-se as métricas área sob a curva (AUC), taxa de falsos positivos ($FPR = \frac{FP}{FP+TN}$), acurácia balanceada ($BACC = \frac{TPR+TNR}{2}$), precisão ($PREC = \frac{TP}{TP+FP}$), *recall* ($REC = \frac{TP}{TP+FN}$) e *f1-score* ($F1 = \frac{PREC \cdot REC}{PREC+REC}$) para a avaliação dos classificadores. Os modelos construídos alcançaram resultados aproximados para as métricas de desempenho consideradas. O modelo M1 foi o que obteve melhor resultado para a métrica AUC, e reflete a aplicação da técnica PCA para redução de dimensionalidade dos dados como passo preliminar ao treinamento da rede neural.

Tabela 1. Avaliação de Desempenho dos Modelos de Redes Neurais MLP

	ATRIB	CAM	AUC	FPR	BACC	PREC	REC	F1
M1	**205 (PCA)**	**1L**	**0.916623**	**0.047009**	**0.777061**	**0.396172**	**0.601131**	**0.477591**
M2	624 (PCA)	2L	0.910612	0.056793	0.789902	0.365124	0.636597	0.464075
M3	624 (PCA)	1L	0.906286	0.067540	0.809330	0.342659	0.686199	0.457074
M4	205 (PCA)	2L	0.905127	0.045321	0.745009	0.377355	0.535338	0.442673
M5	2048	1L	0.899920	0.072946	0.790949	0.315347	0.654845	0.425695
M6	2048	2L	0.897935	0.067395	0.814928	0.346754	0.697250	0.463167

O desempenho da abordagem proposta foi ainda comparado com os trabalhos de [Wan et al. 2020] e [Kamoona et al. 2020], conforme a Tabela 2. Pode-se verificar que o modelo $M1$ alcança resultados comparáveis com o estado da arte em termos de AUC. Todavia, apresenta uma taxa de falsos alarmes alta, quando comparado à abordagem da literatura. Uma explicação possível é que os modelos da literatura são de alta complexidade e assim capazes de realizar tarefas de detecção mais complexas. Enquanto que a rede apresentada em [Wan et al. 2020] faz uso de duas funções de custo para a aprendizagem da anomalia, o modelo proposto é treinado a partir de uma única função de custo (*binary cross entropy*).

Tabela 2. Comparação dos Resultados

Abordagem	AUC (%)	FAR (%)
[Wan et al. 2020] (MIL Loss + Center Loss)	91.24	0.10
[Wan et al. 2020] (MIL Loss)	89.10	0.21
[Kamoona et al. 2020]	89.67	-
Abordagem Proposta (M1)	**91.66**	**4.70**

O gráfico ROC (*Receiver Operating Characteristics*) [Haykin 2010] é uma métrica útil para comparar o desempenho entre diferentes algoritmos e que apresenta o balanceamento entre a *Taxa de Falsos Positivos* (TFP) e a *Taxa de Verdadeiros Positivos* (TVP). O desempenho de um dado modelo de classificação é representado por um

ponto no espaço bidimensional do gráfico ROC. Na Figura 1, são apresentadas as curvas ROC para cada modelo avaliado. Este é um método que permite acomodar a incerteza

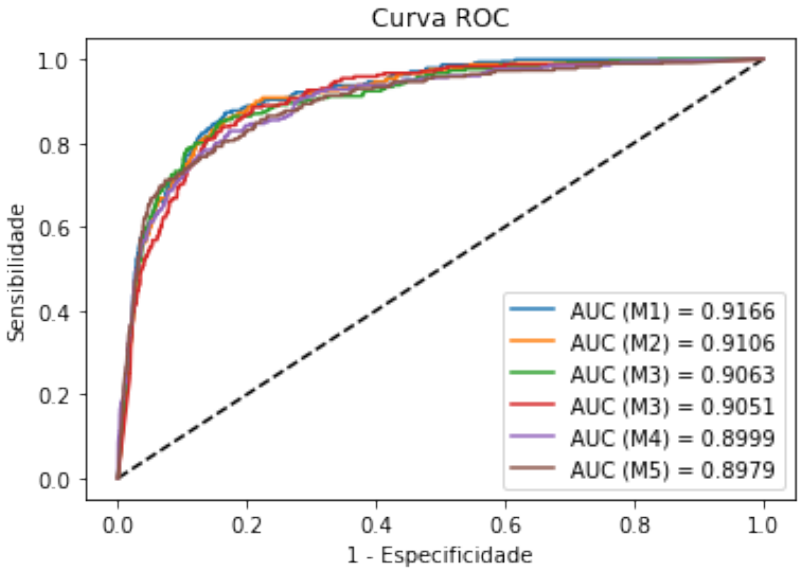

Figura 1. Curvas ROC

ao possibilitar visualizar todo o espaço de possibilidades de desempenho dos estimadores [Haykin 2010]. Pode-se verificar que o desempenho dos diferentes modelos são aproximados, e que há regiões da curva onde um dado valor de teste tem maior taxa de verdadeiros positivos e menor taxa de falsos positivos.

5. Conclusões e Trabalhos Futuros

Neste trabalho foi proposta e avaliada uma abordagem baseada em classificação binária com redes MLP, redução de dimensionalidade com PCA e paradigma MIL para detecção de anomalias em vídeovigilância. O *dataset* de *benchmark ShanghaiTech*, processado e representado como atributos I3D, foi utilizado para modelagem e avaliação da abordagem proposta. A aplicação da técnica PCA reduziu o espaço de entrada com 2048 atributos para 624 e 205 componentes principais, às quais explicam uma quantidade significativa da variância dos dados de treinamento. A abordagem proposta obteve resultados superiores aos dois trabalhos comparados, em termos de AUC.

Direcionamentos futuros incluem a avaliação de outras abordagens para representação de atributos e arquiteturas de redes neurais profundas para a tarefa de videovigilância no contexto de aprendizagem fracamente supervisionada. Ademais, planeja-se experimentos adicionais envolvendo outros *datasets* no intuito de mitigar possíveis dificuldades quando se opera com grandes volumes de dados.

Referências

Ali, S. and Shah, M. (2008). Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 32(2):288–303.

Haykin, S. (2010). *Neural networks and learning machines, 3/E*. Pearson Education India.

Kamoona, A. M., Gosta, A. K., Bab-Hadiashar, A., and Hoseinnezhad, R. (2020). Multiple instance-based video anomaly detection using deep temporal encoding-decoding. *arXiv preprint arXiv:2007.01548*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Li, T., Wang, Z., Liu, S., and Lin, W.-Y. (2021). Deep unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3636–3645.

Nayak, R., Pati, U. C., and Das, S. K. (2020). A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, page 104078.

Pawar, K. and Attar, V. (2019). Deep learning approaches for video-based anomalous activity detection. *World Wide Web*, 22(2):571–601.

Pereira, S. S. L. and Maia, J. E. (2021). Anomaly detection in surveillance video of natural environment. *International Journal of Computer Applications*, 183(1):1–7.

Rao, T. N., Girish, G., and Rajan, J. (2017). An improved contextual information based approach for anomaly detection via adaptive inference for surveillance application. In *Proceedings of International Conference on Computer Vision and Image Processing*, pages 133–147. Springer.

Ribeiro, M., Lazzaretti, A. E., and Lopes, H. S. (2018). A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters*, 105:13–22.

Saha, B. N., Ray, N., and Zhang, H. (2009). Snake validation: A pca-based outlier detection method. *IEEE Signal Processing Letters*, 16(6):549–552.

Suarez, J. J. P. and Naval Jr, P. C. (2020). A survey on deep learning techniques for video anomaly detection. *arXiv preprint arXiv:2009.14146*.

Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488.

Wan, B., Fang, Y., Xia, X., and Mei, J. (2020). Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., and Samaras, D. (2012). Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35. IEEE.

Zhao, H., Lai, Z., Leung, H., and Zhang, X. (2020). *Feature Learning and Understanding: Algorithms and Applications*. Springer Nature.