

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

João Vitor Afonso Pereira

**Técnica de Amostragem de quadros em Vídeos
de Vigilância para Análise Visual de Eventos**

Uberlândia, Brasil

2024

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

João Vitor Afonso Pereira

**Técnica de Amostragem de quadros em Vídeos de
Vigilância para Análise Visual de Eventos**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção título de Ba-
charel em Ciência da Computação.

Orientador: Professor José Gustavo Paiva

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2024

Resumo

Este trabalho introduz uma nova técnica de amostragem de quadros para vídeos de vigilância, visando melhorar a análise visual e compreensão de eventos. A abordagem criada permite a preservação da estrutura original do vídeo por meio de um subconjunto representativo de quadros. Ao mesmo tempo, o número de quadros é significativamente reduzido. O objetivo é discutir a eficácia da abordagem sugerida usando-a em vários cenários de monitoramento, examinando as deficiências dos métodos anteriores e encontrando métodos de amostragem eficazes. Para compor a amostra, os quadros chave de cada grupo foram escolhidos por similaridade na formação dos grupos. Os resultados mostram que a técnica é efetiva na captura da totalidade dos eventos. Em comparação com métodos convencionais como amostragem aleatória e por segundo, ela fornece uma representação mais homogênea e completa. A contribuição desta pesquisa está relacionada ao campo de processamento de vídeos e análise visual, oferecendo uma abordagem adaptável que pode reduzir o volume de dados analisados sem comprometer a integridade dos eventos importantes. Para verificar a eficácia e adaptabilidade da técnica de extração de características, é recomendável estudar várias técnicas e adaptá-las a outros contextos de segurança.

Palavras-chave: Amostragem de quadros, vídeos de vigilância, análise visual, técnicas de agrupamento, representatividade de eventos.

Lista de ilustrações

Figura 1 – Imagem representativa do processo de amostragem de dados de um determinado conjunto.	10
Figura 2 – Imagem representativa do processo de reconhecimento de objetos utilizando uma <i>CNN</i> alterada.(KABANI; EL-SAKKA, 2016)	13
Figura 3 – Múltiplas <i>CNNs</i> para o rastreamento de objetos em vídeo.(LI et al., 2014)	14
Figura 4 – Uso de Multi-Agentes para amostragem de quadros.(WU et al., 2019) .	15
Figura 5 – Diagrama demonstrando a ideia geral da técnica proposta.	17
Figura 6 – Diagrama demonstrando a interação entre as classes do sistema computacional.	21
Figura 7 – Alguns quadros do vídeo Sofa	23
Figura 8 – Evento do carro estacionado saindo do estacionamento do vídeo VIRAT . .	24
Figura 9 – Visualizações do vídeo Sofa completo utilizando as três técnicas extratoras.	25
Figura 10 – Visualizações do vídeo VIRAT completo utilizando as três técnicas extratoras.	25
Figura 11 – Curvas Tempo X Taxa de Similaridade da criação das amostras dos vídeos Sofa e VIRAT	27
Figura 12 – Curvas Tempo X Porcentagem de seleção da amostra dos vídeos Sofa e VIRAT	27
Figura 13 – <i>Layouts</i> das amostras dos vídeos Sofa e VIRAT utilizando uma taxa de similaridade igual a 30%.	29
Figura 14 – <i>Layouts</i> das amostras dos vídeos Sofa e VIRAT utilizando uma taxa de similaridade igual a 50%.	30
Figura 15 – <i>Layouts</i> das amostras dos vídeos Sofa e VIRAT utilizando uma taxa de similaridade igual a 90%	31
Figura 16 – <i>Layouts</i> das amostras dos vídeos Sofa e VIRAT para variação da porcentagem de quadros da amostra utilizando HOG	33
Figura 17 – <i>Layouts</i> das amostras dos vídeos Sofa e VIRAT para variação da porcentagem de quadros da amostra utilizando CNN	34
Figura 18 – <i>Layouts</i> dos vídeos Sofa e VIRAT completos com HOG e CNN	36
Figura 19 – <i>Layouts</i> das amostras da técnica proposta dos vídeos Sofa e VIRAT para HOG e CNN	37
Figura 20 – <i>Layouts</i> das amostras dos vídeos Sofa e VIRAT utilizando amostragem aleatória e amostragem por segundo	38

Lista de tabelas

Tabela 1	– Tabela de eventos do vídeo Sofa	24
Tabela 2	– Tabela de eventos relacionados a um automóvel saindo do estacionamento do vídeo VIRAT	24

Sumário

1	INTRODUÇÃO	7
1.1	Objetivos	8
1.2	Justificativa	8
1.3	Organização do texto	8
2	REFERENCIAL TEÓRICO	9
2.1	Terminologia	9
2.2	Amostragem de Dados	10
2.3	Visualização de Informação	11
2.3.1	Projeções Multidimensionais de Dados	12
2.4	Trabalhos Relacionados	12
2.4.1	Considerações finais	16
3	METODOLOGIA	17
3.1	Descrição da Técnica	17
3.1.1	Módulo de extração de quadros e características	18
3.1.2	Módulo de agrupamento de quadros	18
3.1.3	Módulo de amostragem	20
3.2	Descrição do Sistema Computacional	20
3.2.1	Infraestrutura Computacional	20
3.2.2	Implementação do Algoritmo de Agrupamento	21
4	EXPERIMENTOS E RESULTADOS	22
4.1	<i>Visual Classification System (VCS)</i>	22
4.2	Vídeos	23
4.2.1	Sofa	23
4.2.2	VIRAT	23
4.3	Experimentos	24
4.3.1	Avaliação das Técnicas de Extração de características	25
4.3.2	Análise Computacional	26
4.3.3	Experimento de variação da taxa de similaridade	28
4.3.4	Experimento de variação da porcentagem de quadros da amostra	32
4.3.5	Análise da Representatividade Amostral	35
5	CONCLUSÃO	40

REFERÊNCIAS 41

1 Introdução

Na atualidade a segurança é um tema de significativa importância tanto em âmbito pessoal quanto social. Com o avanço tecnológico as câmeras se tornaram populares na área de segurança, produzindo grande quantidade de material na forma de arquivos de vídeo. Identificar acontecimentos estratégicos no conteúdo desses vídeos, tais como assaltos, acidentes ou padrões de movimentação é essencial e justifica a necessidade de uma análise eficaz dos dados. No entanto, a análise manual desse material de vídeo é trabalhosa e pouco eficiente, pois em sua maioria são vídeos de longa duração que exigem um alto esforço humano de análise. Além disso, muitos eventos podem ocorrer simultaneamente no vídeo, e há o risco de o resultado da análise humana ser incompleta dada a realidade das ações captadas.

Existem diversas técnicas computacionais que auxiliam na análise do material capturado, realizando procedimentos automáticos para a tomada de decisão, que se utilizam de conceitos relacionados ao aprendizado de máquinas de forma a extrair bons resultados. As técnicas de visualização de informação, por outro lado, têm como objetivo criar representações gráficas que realcem padrões e tendências que explicam o comportamento dos dados em questão. Tais técnicas possibilitam ressaltar ao profissional de segurança aspectos importantes do vídeo para a análise, como trajetórias, ocorrência de eventos, comportamento dos objetos em cena, entre outras.

Apesar de obterem resultados satisfatórios em diversas tarefas de análise dos vídeos, grande parte dessas estratégias visuais opera em todos os quadros do vídeo, um a um, o que faz com que essas técnicas sejam computacionalmente caras, e gere *layouts* de difícil interação. Considerando que em muitas situações são utilizadas câmeras fixas, é alta a chance de que em uma considerável porção do vídeo não haja movimentos, ou haja pequenos movimentos, e, portanto, os quadros frequentemente não sofrem alterações significativas. Existem diversas estratégias de amostragem de dados documentadas na literatura, que aplicam diferentes abordagens para extrair um subconjunto dos dados que represente bem o seu conteúdo total em termos de fenômenos e comportamentos. Para o contexto de vigilância apresentado anteriormente, é possível que tais técnicas possam ser aplicadas aos vídeos, no sentido de obter um subconjunto de quadros que represente adequadamente os acontecimentos do vídeo, diminuindo o custo das técnicas tradicionalmente empregadas nessas análises.

1.1 Objetivos

Este trabalho pretende propor uma nova técnica amostragem em vídeos de monitoramento, com foco na análise visual para identificação e compreensão de eventos ocorridos. A ideia é desenvolver uma técnica capaz de gerar um subconjunto de quadros que representem de forma satisfatória os eventos contidos nos vídeos, de forma a manter a estrutura original do vídeo com uma quantidade reduzida de quadros.

Pretende-se discutir os resultados obtidos a partir da implementação da técnica proposta e aplicação em vídeos de monitoramento que representem situações diversas, de forma avaliar a capacidade de captura dessas situações.

Ao longo do trabalho tem-se os seguintes objetivos específicos:

- Identificar potenciais estratégias de amostragem utilizada, bem como limitações das técnicas existentes;
- Desenvolver uma técnica de amostragem de quadros em vídeos de vigilância que capture os diferentes eventos ocorridos;
- Desenvolver um sistema que incorpora a técnica desenvolvida;
- Ilustrar a aplicação da técnica proposta em vídeos que representem diversos cenários relacionados à vigilância.

1.2 Justificativa

Avaliar as informações captadas em vídeos de vigilância é fundamental para auxiliar a tomada de decisão dos profissionais de segurança em diversos cenários. Diversas estratégias computacionais existem para facilitar na análise do material disponível, e acredita-se que a utilização de técnicas de amostragem podem beneficiar essas estratégias, no sentido de encontrar um subconjunto do material de vigilância que preserve o conteúdo interessante de análise e ao mesmo tempo diminua o custo computacional da geração dos *layouts*, tornando-as aptas para aplicação em um número maior de cenários.

1.3 Organização do texto

Nas seções posteriores são descritos o referencial teórico, trabalhos relacionados, a técnica proposta para a amostragem de quadros de um vídeo de segurança, bem como o sistema computacional que incorpora a técnica e os resultados do uso da técnica em experimentos de amostragem de quadros.

2 Referencial Teórico

Neste capítulo serão apresentados termos e conceitos relacionados a amostragem de quadros de vídeos de segurança, os quais serão utilizados na apresentação dos trabalhos correlatos, bem como na descrição da proposta e apresentação dos resultados obtidos.

2.1 Terminologia

- Quadro: Imagem individual inserida em uma sequência de imagens que formam um vídeo;
- Vídeo: Coleção de quadros organizados e exibidos em sequência, que representam algum tipo de fenômeno ocorrido ao longo de um período de tempo;
- Amostragem: Processo que por meio de técnicas de seleção escolhe membros de um conjunto de dados de forma que por meio de uma análise dos selecionados, seja possível realizar inferências sobre o conjunto de dados completos. Este termo será explorado em detalhes na próxima subseção;
- Taxa de quadros: No cenário de vídeos de segurança, é a quantidade de quadros presentes em um espaço temporal específico, geralmente 1 segundo. Assim, um vídeo que possui 24 quadros por segundo significa que em um único segundo estão inseridos 24 quadros individuais que em conjunto criam a ilusão de movimento dentro desse espaço de tempo. Quanto maior o número de quadros em um determinado período de tempo mais suave o movimento parecerá;
- Evento: Acontecimento ou fenômeno observável em um período de tempo captado por câmeras e que potencialmente contém informação relevante para o contexto de segurança;
- Cena: Local físico captado em vídeo, ou região da imagem no qual ocorre um evento de interesse;
- Vigilância: Atividade de monitoramento e acompanhamento de fenômenos que ocorrem em determinado local, normalmente com o objetivo de garantir a aplicação da lei, segurança pública, individual ou patrimonial;
- Vídeo de Segurança ou Vídeo de monitoramento: Vídeo produzido por câmeras de segurança, fixas ou móveis, posicionadas em locais estratégicos para garantir a segurança de um bem público ou privado, ou de indivíduos.

2.2 Amostragem de Dados

O conceito de amostragem de dados consiste em técnicas de análises estatísticas utilizadas para selecionar, manipular e analisar um subconjunto de todos os dados coletados, com objetivo de descobrir informações relevantes no conjunto de dados originais (BOLFARINE; BUSSAB, 2005). A utilização dessas técnicas são particularmente úteis ao se trabalhar com um conjunto de dados extenso (ACHARYA et al., 2013), uma vez que, identificar e analisar uma amostra representativa é mais eficiente e menos custoso do que realizar a análise no conjunto completo.

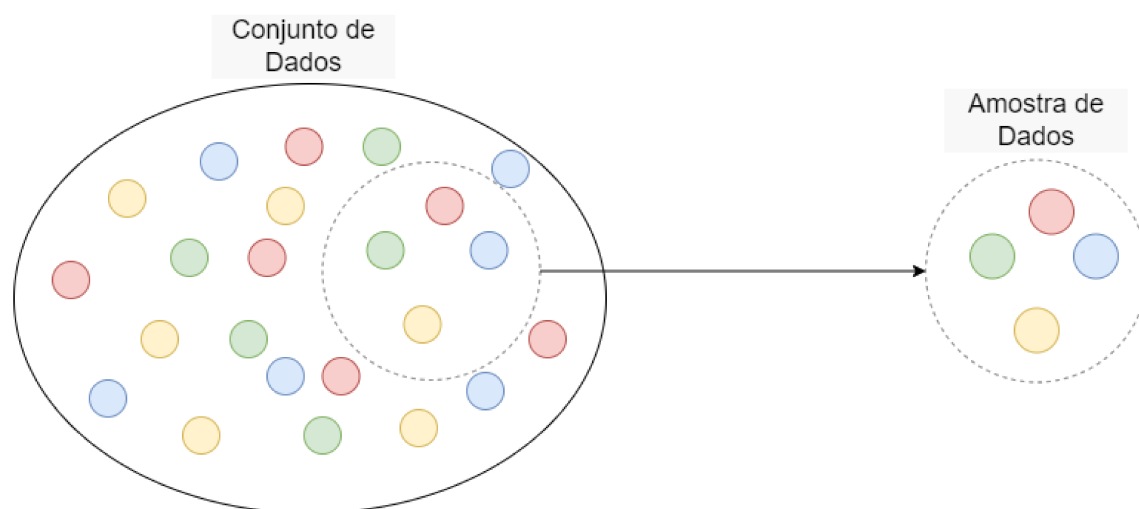


Figura 1 – Imagem representativa do processo de amostragem de dados de um determinado conjunto.

Uma consideração importante para a amostragem de dados é determinar o tamanho ideal da amostra bem como a técnica implementada para gerar essa amostra, visto que o uso inadequado de uma técnica de amostragem pode gerar um viés de interpretação dos resultados (BOLFARINE; BUSSAB, 2005), ou gerar um conjunto no qual parte dos fenômenos não ocorre.

Existem diferentes técnicas para a construção de uma amostra representativa, dependendo dos dados que estão sendo estudados. De forma geral, as técnicas dividem-se em técnicas de amostragem probabilísticas e não probabilísticas. A amostragem probabilística permite inferir as descobertas sobre a amostra para o conjunto de dados original (ACHARYA et al., 2013), pelo fato de cada integrante dos dados originais possuem a mesma probabilidade de serem escolhidos para constituir a amostra. Assim, o conjunto amostral pode ser representativo de todas as informações relevantes dos dados originais. Já na amostragem não probabilística a seleção dos dados que compõem a amostra é feita a partir da escolha do próprio pesquisador, ou selecionadas a partir de uma cota pré-estabelecida (ACHARYA et al., 2013), como por exemplo selecionar um número X de quadros que apresentam uma árvore. Dessa maneira, amostras não probabilísticas são mais propensas a resultar em um viés de interpretação, uma vez que nem todos os dados

originais possuem a mesma probabilidade de compor a amostra. Assim, as conclusões alcançadas pela análise da amostra podem não ser representativas do conjunto original já que características importantes podem não ter sido selecionadas.

A aplicação dos conceitos de amostragem de dados no cenário específico em que o conjunto original de dados é um vídeo ou gravação é denominada de amostragem de quadros. Nesse caso, o objetivo é obter um conjunto de quadros que represente adequadamente algum aspecto de análise, por exemplo, dos eventos ocorridos nesse vídeo. O uso de técnicas de amostragem de quadros se faz essencial, porque é comum que as gravações tenham muitas horas de duração, e resultem em arquivos grandes, o que gera problemas de armazenamento. Além disso, técnicas computacionais de análise desses vídeos geralmente possuem um custo de execução alto, e diminuir o conjunto no qual elas operam pode trazer benefícios para a análise.

2.3 Visualização de Informação

A área de Visualização de Informação estuda maneiras de representar dados de uma forma visual expressiva, com o intuito de facilitar a interpretação e a propiciar uma análise mais efetiva desses dados, sendo um meio eficaz no compartilhamento de informações de modo natural para alguém não especializado no tópico. Além disso, essas técnicas são capazes de apresentar relações e conexões relevantes nos dados permitindo-se chegar a conclusões de forma mais simples. Logo Visualização de Informação é utilizar representações visuais iterativas de dados abstratos de forma a ampliar o entendimento sobre eles.

Existem várias técnicas de Visualização de Informação, dentre elas tem-se algumas mais simples e amplamente utilizadas não somente no desenvolvimento acadêmico, mas também disseminadas em uso popular ([SADIKU et al., 2016](#)), como o gráfico de colunas, gráfico de barras, mapa hexadecimal, histograma, entre outros.

Neste trabalho será utilizado o sistema visual produzido no trabalho ([PAIVA et al., 2014](#)) como sistema base no qual serão visualizadas as amostras criadas pela técnica proposta, detalhes sobre esse sistema visual serão apresentados em uma seção posterior.

No contexto de amostragem de quadros em vídeos de segurança, o uso de técnicas de Visualização de Informação como as que serão apresentadas durante o desenvolvimento deste trabalho é de fundamental importância, uma vez que possibilita a representação visual da qualidade da amostragem realizada a partir da representação tanto espacial quanto por gradação de cor da similaridade das informações contidas em cada quadro que compõem a amostra.

2.3.1 Projeções Multidimensionais de Dados

Dados multidimensionais são um conjunto de dados que possuem diferentes características e dados complexos como um quadro de uma gravação de segurança normalmente possuem várias características diferenciadoras, como gradação de cores, movimentação de objetos e sons, as quais também são denominadas de variáveis, dimensões e instâncias, (ESPADOTO; HIRATA; TELEA, 2020), quanto maior o número de dimensões de um conjunto de dados maior o número de informações contido no mesmo. Assim para extrair e analisar tais informações é realizada uma projeção dessas informações multidimensionais dos dados, as quais utilizam de técnicas de redução de dimensionalidade, da original para 2 ou 3 dimensões, com o objetivo de produzir um gráfico de dispersão para análise visual.

Estas projeções mapeiam dados de um espaço multidimensional para um espaço visual, mapeando as instâncias ou características ou dimensões dos dados coletados em um espaço geralmente bidimensional, preservando as relações de similaridade entre os dados projetados (NEVES, 2016). Os gráficos de dispersão resultantes são caracterizados por conjuntos de pontos posicionados no plano, em que cada ponto corresponde a uma instância no conjunto original de dados. Em uma boa projeção, pontos projetados próximos indicam características similares (NEVES, 2016), ou seja, dados com conteúdo similar e pontos distantes indicam dados com conteúdo dissimilares, de acordo com algum aspecto. No contexto de vídeos de segurança, é comum que as técnicas representem o vídeo como uma coleção de seus quadros constituintes. Dessa maneira, o resultado da aplicação de uma técnica de projeção multidimensional exibirá quadros com conteúdo similar em regiões próximas do *Layout*, e quadros com conteúdo dissimilar em regiões distantes entre si.

2.4 Trabalhos Relacionados

Nesta seção serão apresentados diversos trabalhos relacionados à proposta descrita neste trabalho, de forma a analisar seus benefícios e limitações, e como podem de alguma forma contribuir para o desenvolvimento deste trabalho. Além de serem apresentadas publicações que se utilizam de técnicas relevantes à abordagem apresentada posteriormente.

No contexto de vídeos normalmente os dados a serem amostrados são representados pelos quadros que compõem o vídeo, e o resultado será um subconjunto dos quadros totais originais. O objetivo de forma abrangente é gerar uma amostra composta pelos quadros mais significativos, que contenham os eventos mais relevantes para a análise em questão.

A pesquisa e desenvolvimento de métodos e técnicas de amostragem de quadros vem adquirindo grande atenção, devido ao uso disseminado de vídeos em diversas áreas, como recomendação de vídeos, segurança, reconhecimento de vídeos, entre outros. Grande parte das técnicas de amostragem de quadros de um vídeo se utilizam de três conceitos

de reconhecimento de características presentes em um quadro, (*Histogram of Oriented Gradient* (HOG)), de aprendizado de máquinas, em virtude da capacidade de aprendizado e melhora na qualidade da amostragem a partir de um número de iterações.

O aprendizado de máquina é uma abordagem multidisciplinar com o objetivo de criar técnicas computacionais que detectam padrões e fenômenos em coleções de dados automaticamente (CAMASTRA; VINCIARELLI, 2015).

A análise automática de vídeos usando aprendizado de máquinas pode ser realizada através de diversas técnicas. Dentre elas, destaca-se o uso de *Convolutional Neural Networks* (CNN), que representam técnicas de aprendizagem profunda que recebe imagens como entradas, aplica pesos (medidas matemáticas de importância) para vários aspectos das imagens e extrair essas informações abstratas. Assim, as CNNs são capazes de representar informações visuais de forma distribuída, e avaliar uma grande quantidade de dados (SONG et al., 2019), normalmente sendo aplicada para a seleção de quadros significativos.

O trabalho proposto por (KABANI; EL-SAKKA, 2016) propõe o uso de uma CNN para o reconhecimento de objetos presentes em quadros específicos do vídeo, como mostrado na Figura 2.

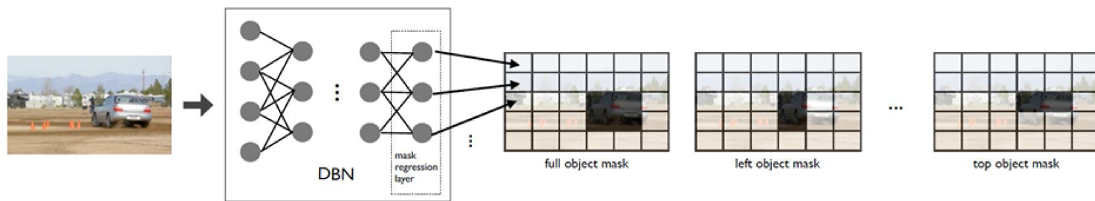


Figura 2 – Imagem representativa do processo de reconhecimento de objetos utilizando uma CNN alterada. (KABANI; EL-SAKKA, 2016)

Redes neurais convolucionais também podem ser utilizadas para realizar o rastreamento de um objeto identificado em um vídeo, como proposto por (LI et al., 2014), que utilizou características discriminativas dos objetos para o seu rastreo. Diversas CNNs são utilizadas para manter diversos núcleos com diferentes características simples, para diferenciar o objeto do ambiente em que se encontra, como evidenciado na Figura 3

Alguns trabalhos como (TRAN et al., 2015), (CARREIRA; ZISSERMAN, 2017) e (FEICHTENHOFER, 2020) utilizam redes neurais 3D (*3D Convolutional Neural Networks*, ou 3D CNN) para realizar a amostragem de quadros de um vídeo completo, em que um quadro aleatório é escolhido como ponto de partida. Os 64 quadros consecutivos juntamente com o quadro inicial são então agrupados em uma amostra, assim a partir do momento em que a rede neural é treinada em tal processo a escolha de quadros iniciais converge para quadros significativos do vídeo, ou seja, quadros com acontecimentos relevantes.

Já o trabalho proposto por (WANG et al., 2016) utiliza uma Rede de segmento

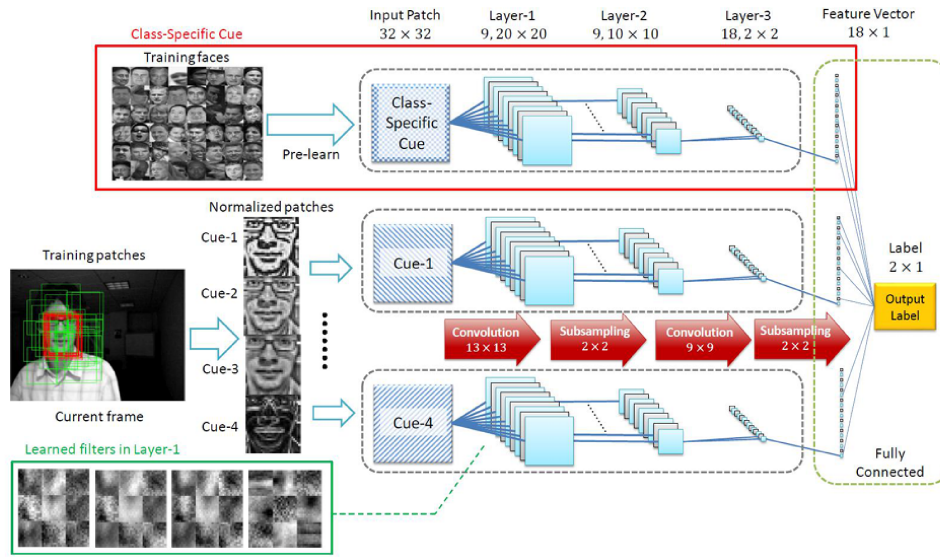


Figura 3 – Múltiplas CNNs para o rastreamento de objetos em vídeo. (LI et al., 2014)

temporal (*Temporal Segment Network*, ou TSN) para realizar o processo de amostragem de quadros de um vídeo. Nessa estratégia um vídeo é dividido em K segmentos de tamanho temporal uniforme, por exemplo segmentos compostos pelos quadros presentes em 5 segundos consecutivos. Após essa segmentação um conjunto de quadros de cada segmento é escolhido aleatoriamente, sendo analisado como uma amostra. Então, a TSN é usada para extrapolar as informações dessa amostra para toda a duração do vídeo, compondo o conjunto de amostras final. Assim, essa estratégia é capaz de analisar um pequeno grupo de quadros do vídeo em tempos diferentes e por meio da rede de segmento temporal gerar amostras significativas.

As estratégias apresentadas anteriormente tratam cada quadro de forma igualitária, visto que utilizando 3D CNN há a probabilidade de um quadro não significativo ser escolhido como ponto de partida. Assim as ambas as técnicas ignoram a redundância de informações contidas nos quadros, ou seja, amostras diferentes podem possuir quadros com as mesmas informações.

Como alternativa alguns trabalhos utilizam redes neurais com aprendizado reforçado (*Reinforced Learning*, ou RL). O trabalho (FAN et al., 2018) propõe um software chamado *FastForward* que utiliza RL tanto para o planejamento de escolha de quadros, como também para a decisão prévia do ponto de parada pra o fim da amostra, com o intuito de reduzir o custo computacional de se amostrar todos os quadros de vídeos longos, já que em sua maioria não são significativos para uma análise de amostragem específica.

Outra abordagem presente em (WU et al., 2019) utiliza multiagentes em uma rede de observação sensível ao contexto, em que N agentes observadores de forma individual são responsáveis por selecionar um quadro significativo cada, criando uma amostra composta somente por quadros relevantes, cada N agente é representado por uma CNN que reali-

zam a seleção dos quadros significativos, inicialmente tal processo é realizado de forma aleatória, mas com as iterações e mudanças de pesos da rede de acordo com significância dos quadros escolhidos a seleção torna-se assertiva com o tempo. Assim, a cada nova amostragem realizada, os agentes são treinados e passam a escolher quadros com uma taxa de significância maior, melhorando a qualidade da amostra gerada como resultado, como apresentado na Figura 4.

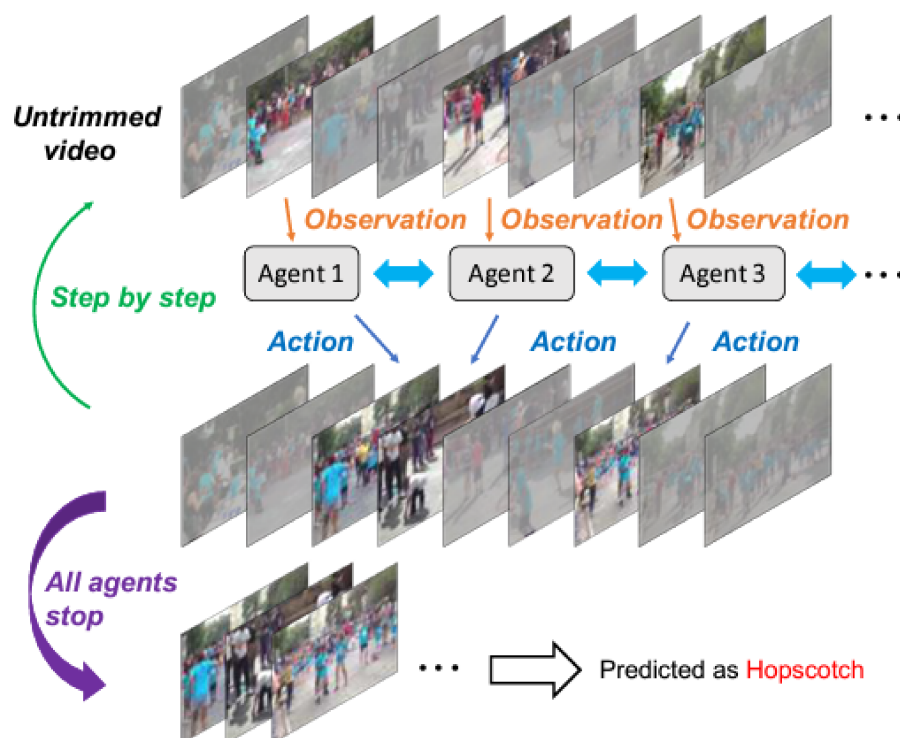


Figura 4 – Uso de Multi-Agentes para amostragem de quadros.(WU et al., 2019)

O trabalho (ZHI et al., 2021) propõe uma Interface de Programação de Aplicação (API) chamada MGSampler, o qual é uma estratégia de amostragem para análise de vídeos que utiliza amostragem aleatória guiada por informações de movimento. Ele seleciona um subconjunto de quadros de um vídeo completo de forma adaptativa, com base na distribuição cumulativa da magnitude do movimento ao longo do tempo. Isso permite uma representação compacta e eficiente do vídeo, capturando os segmentos mais relevantes para o reconhecimento de ações. A abordagem é projetada para ser geral e flexível, podendo ser aplicada a qualquer rede de reconhecimento de vídeo.

Em (DAS; SARKER; MAHMUD, 2019) e (BURGOS-ARTIZZU et al., 2012) é utilizado a técnica HOG para a extração das características dos quadros do vídeo para sua posterior coleta para a amostra. O HOG é uma técnica de extração de características utilizada em visão computacional e processamento de imagens, popularizada por Navneet Dalal e Bill Triggs (DALAL; TRIGGS, 2005). Essa técnica é eficaz na identificação de formas e contornos de objetos ao analisar as direções das bordas na imagem. O HOG é amplamente utilizado em diversas aplicações de visão computacional, como detecção de

pedestres, reconhecimento de rostos e classificação de objetos, devido à sua capacidade de capturar características essenciais sem depender de informações de cor.

2.4.1 Considerações finais

Nesta seção, discutiram-se técnicas de amostragem de quadros em vídeos, como redes neurais convolucionais (CNNs), redes neurais 3D (3D CNNs), Redes de Segmento Temporal (TSNs), aprendizado reforçado (RL) e a técnica de Histograma de Gradientes Orientados (HOG), cada uma com seus benefícios e limitações. CNNs e 3D CNNs são utilizadas para seleção de quadros significativos e rastreamento de objetos, mas podem ser computacionalmente intensivas. TSNs analisam pequenos grupos de quadros, mas podem não capturar completamente a dinâmica temporal de eventos importantes. RL e multia-gentes reduzem o custo computacional na seleção de quadros relevantes, mas enfrentam desafios de convergência e treinamento. HOG é eficaz na extração de características, mas limitado em cenários onde cor e textura são cruciais. Desafios em aberto incluem equilibrar eficiência computacional e precisão na seleção de quadros, reduzir redundância de informações entre quadros consecutivos e adaptar técnicas a diferentes contextos de vídeos.

3 Metodologia

Neste capítulo a técnica de amostragem de quadros é detalhada, assim como, o sistema computacional que a implementa.

3.1 Descrição da Técnica

A técnica proposta neste trabalho tem como objetivo gerar uma amostra composta por quadros representativos dos diferentes eventos presentes em um vídeo de vigilância, para ser utilizada em diversas tarefas de análise. A Figura 5 apresenta o diagrama de funcionamento da técnica, cujos passos são detalhados a seguir.

A técnica possui três módulos distintos: um módulo de extração de quadros e características, implementando três técnicas de extração de características de quadros distintas; um módulo de agrupamento, que utiliza as características já extraídas para a criação de grupos de quadros baseados em sua similaridade e um módulo de amostragem que cria a amostra final selecionando-se um conjunto de quadros de cada grupo para compor a amostra final do tamanho desejado.

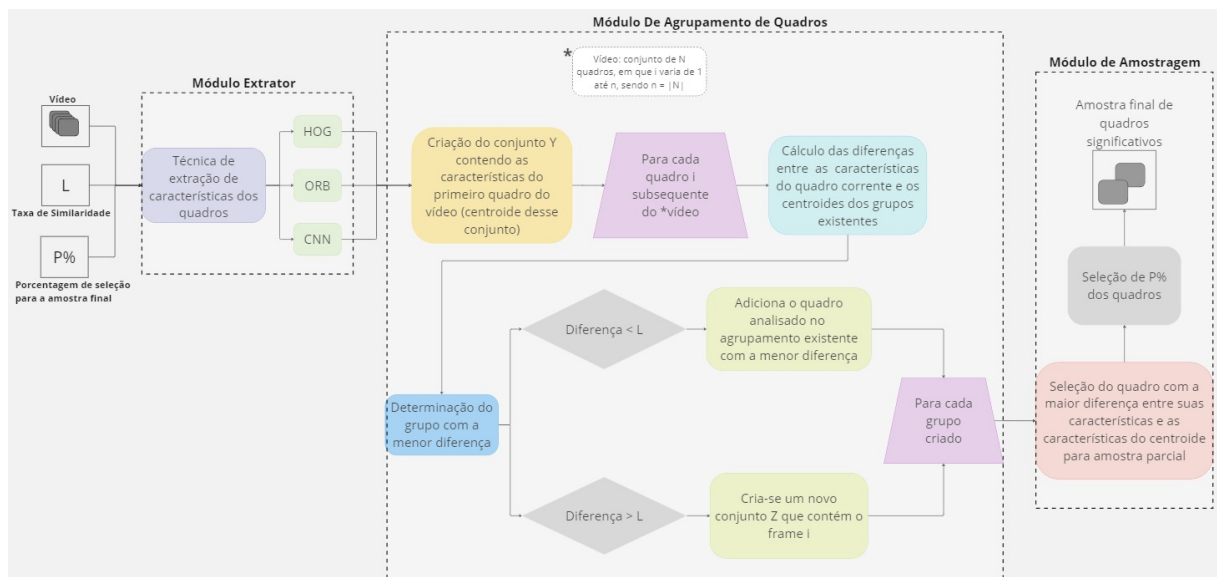


Figura 5 – Diagrama demonstrando a ideia geral da técnica proposta.

Além disso a técnica possui dois parâmetros necessários para seu funcionamento:

- L : taxa de similaridade mínima entre os quadros do vídeo, utilizado para guiar a definição dos eventos em termos de alterações na cena;
- P : porcentagem de quadros, de cada grupo formado, que comporão a amostra final.

O impacto de cada um desses parâmetros são discutidos nas subseções seguintes.

3.1.1 Módulo de extração de quadros e características

Este módulo é responsável por extrair características que descrevam aspectos importantes em cada um dos quadros do vídeo. Foram utilizadas três técnicas de extração distintas: *Histogram of Oriented Gradients (HOG)* (DALAL; TRIGGS, 2005), *Oriented FAST and Rotated BRIEF (ORB)* (RUBLEE et al., 2011) e *Convolutional Neural Networks (CNN)* (SIMONYAN; ZISSERMAN, 2015), descritas a seguir:

- HOG: calcula e contabiliza as direções dos gradientes de intensidade de luz nos quadros do vídeo. O processo de extração de características envolve a divisão da imagem em células pequenas, cálculo de histogramas de gradientes orientados em cada célula, normalização dos histogramas e concatenação para formar o descritor final (DALAL; TRIGGS, 2005). Esse descritor é um vetor de características que representa a aparência e a forma do objeto dentro do quadro. Logo é eficaz para reconhecer formas e contornos, usada na detecção de objetos e pessoas;
- ORB: representa uma combinação do detector de pontos-chave *FAST* (ROSTEN; DRUMMOND, 2006) e do descritor de características *BRIEF* (CALONDER et al., 2010), com algumas modificações para melhorar a performance e a robustez. A técnica detecta os pontos-chave mais salientes e calcula os descritores que são invariantes à rotação e resistentes a ruídos (RUBLEE et al., 2011), eficaz no rastreamento de objetos em tempo real e reconhecimento de ambientes;
- CNN: abordagem baseada em redes neurais profundas que atuam no aprendizado de hierarquias de características de baixo a alto nível, o que as torna particularmente adequadas para tarefas de reconhecimento de padrões complexos. A extração de características geralmente envolve o uso de modelos pré-treinados, como o VGG16 a qual foi usada neste trabalho, e a obtenção de características das camadas intermediárias da rede (SIMONYAN; ZISSERMAN, 2015). Sendo assim capaz de diferenciar alterações sutis, formas e movimentos complexos.

O processo de amostragem utiliza uma dessas abordagens para a extração das características dos quadros que compõe o vídeo, a avaliação de desempenho de cada uma é apresentada na seção 4.3.1

3.1.2 Módulo de agrupamento de quadros

Após a etapa de extração de características o módulo de agrupamento organiza as características extraídas dos quadros do vídeo em grupos significativos de quadros, o

módulo recebe como parâmetros a taxa de similaridade L e a porcentagem de quadros da amostra P representados no diagrama da Figura 5. O objetivo é identificar conjuntos de quadros similares, com base em uma métrica de diferença específica para cada tipo de descritor extraído por cada uma das técnicas, e selecionar quadros representativos para compor a amostra final.

Para cada grupo formado, um centroide é calculado, definido pela média das características dos quadros pertencentes ao grupo. O centroide C de um grupo de quadros Q com descritores D_i é dado por:

$$C = \frac{1}{|Q|} \sum_{i=1}^{|Q|} D_i \quad (3.1)$$

onde $|Q|$ o número total de quadros pertencentes ao grupo, e D_i é o descritor do i -ésimo quadro.

A diferença entre as características do quadro corrente e o centroide do grupo é calculada utilizando a métrica apropriada para o tipo de descritor extraído.

Para descritores HOG e CNN a diferença é calculada utilizando-se a distância Euclidiana como métrica diferenciadora. Sendo D_i o descritor do quadro corrente e C o centroide de grupo existente, a diferença é calculada por:

$$d(D_i, C) = \sqrt{\sum_{j=1}^n (D_{i,j} - C_j)^2} \quad (3.2)$$

onde n é o número de elementos no descritor, que pode variar se é HOG ou CNN.

Já para descritores ORB, que são binários, a distância de Hamming é utilizada para calcular a diferença. Se D_i for o descritor binário do quadro corrente e C o centroide do grupo, a distância de Hamming é dada por:

$$d_{ORB}(D_i, C) = \sum_{j=1}^n D_{i,j} \oplus C_j \quad (3.3)$$

onde $\oplus$ denota a operação *XOR* bit a bit entre os elementos dos descritores, n é o número de bits do descritor ORB, e C_j é o centroide do grupo.

O processo de agrupamento consiste então em iterar sobre cada quadro do vídeo e calcular a diferença entre as suas características e os centroides dos grupos existentes para se determinar o centroide com menor diferença. Dado um limiar L , e a menor diferença $d_{Min}(D_i, C)$:

- Se $d_{Min}(D_i, C) < L$, então o quadro i é agregado ao grupo;
- Se $d_{Min}(D_i, C) \geq L$, então cria-se um novo grupo contendo o quadro i .

3.1.3 Módulo de amostragem

Após a formação dos grupos, dado um grupo com N quadros e um centroide C é selecionada uma amostra parcial de quadros até alcançar o percentual P do total de quadros do grupo. Esta seleção inclui tanto os $\frac{P}{2} \times N$ quadros que têm as maiores diferenças em relação ao centroide C , quanto os $\frac{P}{2} \times N$ quadros que têm as menores diferenças. Dessa maneira, a amostra parcial contém quadros que representam os extremos de variabilidade dentro do grupo, garantindo uma representação equilibrada dos eventos ao longo do vídeo.

Finalmente, a amostra final é formada pela união das amostras parciais de todos os grupos, proporcionando uma representação abrangente e diversificada das características observadas nos diferentes grupos.

3.2 Descrição do Sistema Computacional

O desenvolvimento de uma solução computacional eficaz para o agrupamento e amostragem de quadros de vídeo exige uma abordagem sistemática que considere tanto a precisão algorítmica quanto a eficiência computacional. Esta seção detalha a infraestrutura computacional utilizada e as escolhas de implementação que possibilitam a operacionalização da técnica de amostragem descrita anteriormente.

3.2.1 Infraestrutura Computacional

A implementação do sistema de amostragem de quadros foi realizada utilizando a linguagem de programação Java, juntamente com o uso das bibliotecas JavaCV ([AURET; CONTRIBUTORS, 2024](#)) que é uma interface para a biblioteca OpenCV ([Open Source Vision Foundation, 2024](#)) para processamento de imagens e vídeos, Deeplearning4j ([Eclipse Foundation, 2024](#)) para manipulação de redes neurais convolucionais (CNN), e outras ferramentas auxiliares para a manipulação de dados e otimização do desempenho. O sistema foi desenvolvido e testado em uma máquina com as seguintes especificações:

- Processador: Intel Core i7-9750H CPU @ 2.60GHz
- Memória RAM: 16GB DDR4
- Sistema Operacional: Windows 10 Home, 64-bit
- JDK versão: OpenJDK 17
- JavaCV versão: 1.5.9
- Deeplearning4j versão: 1.0.0-M2.1

3.2.2 Implementação do Algoritmo de Agrupamento

O algoritmo recebe como parâmetros o vídeo completo na formatação *MPEG-4 Part 14 (MP4)* ([International Organization for Standardization, 2020](#)), a taxa de similaridade L , a qual é uma porcentagem de 0% a 100% e a porcentagem de quadros para a amostra final P que também varia entre 0% e 100%. O algoritmo tem como produtos de saída um diretório contendo os quadros pertencentes a amostra, como também a amostra em um arquivo texto com a formatação *Points File (.data)* ([ELER et al., 2008](#)). As principais classes e métodos responsáveis pela execução do processo de agrupamento são descritas a seguir:

- **FrameExtractor**: Classe responsável pela coleta dos quadros do vídeo e acionamento das classes apropriadas para a extração das características correspondentes, além de armazená-las em um arquivo texto.
- **HOGExtractor**, **ORBExtractor**, **CNNExtractor**: Classes concretas acionadas pela classe *FrameExtractor* para cada técnica de extração de características.
- **FrameService**: Classe encarregada de gerenciar o processo de agrupamento, calculando diferenças, formando grupos e selecionando quadros representativos para a amostra final que recebe as características dos quadros já extraídas.
- **PointsFileGenerator**: Classe encarregada de criar o diretório da amostra contendo os quadros representados na mesma, bem como criar e salvar o arquivo *Points File (.data)* referente a amostra, acionada pela classe *FrameService*.

A interação entre essas classes, bem como os parâmetros de cada uma, bem como a formatação da amostra final são apresentados na Figura 6.

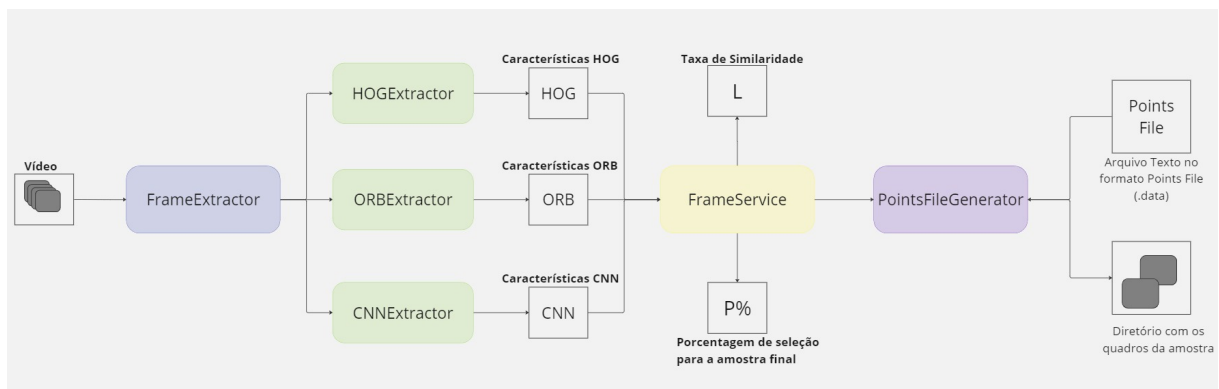


Figura 6 – Diagrama demonstrando a interação entre as classes do sistema computacional.

4 Experimentos e Resultados

Nesta seção, são apresentados detalhes dos vídeos e dos experimentos realizados para avaliar a eficácia da técnica de Amostragem de Quadros em vídeos de segurança descrita na Seção 3.1. Os experimentos foram projetados para analisar diversos aspectos da técnica, incluindo o impacto de configurações do próprio algoritmo, a eficiência computacional e a capacidade das amostras de representar todos os eventos presentes no vídeo, bem como representar aspectos mínimos da estrutura básica de cada evento. A seguir são apresentados os vídeos utilizados, bem como a plataforma de visualização utilizada nos experimentos de avaliação das amostras produzidas.

4.1 *Visual Classification System (VCS)*

O sistema de classificação visual proposto no trabalho (PAIVA et al., 2014) ou *VCS* é uma ferramenta computacional que emprega uma metodologia de Visualização da Informação. Essa metodologia emprega técnicas de visualização baseadas no posicionamento de pontos para construir *layouts* que ajudam o usuário a visualizar e ajustar conjuntos de dados utilizados para treinamento, construir e aplicar modelos de classificação usando diversas técnicas presentes na literatura, bem como realizar uma análise profunda dos resultados da classificação. No trabalho foram utilizadas apenas as funcionalidades referentes a visualização de dados.

O sistema processa os dados a partir de um formato específico chamado *Points File (.data)* (ELER et al., 2009). Este formato de arquivo é utilizado para representar o espaço multidimensional das características de uma imagem ou quadro. O arquivo é organizado em linhas, onde cada uma corresponde a um quadro. Existem duas formas de representação nesse formato: a densa e a esparsa. Neste trabalho, optou-se pela representação densa, na qual cada linha começa com um identificador único e é seguida pelas coordenadas do vetor de características, sendo cada elemento separado por ponto e vírgula.

Adicionalmente, o usuário pode também através dos *layouts* criados, ajustar os modelos de forma a adaptá-los a cenários de classificação específicos. Além das tarefas relacionadas ao processo de classificação, o sistema também oferece diversas funcionalidades para explorar os *layouts* criados, a fim de detectar particularidades e tendências no conjunto de dados que ajudam a entender as decisões dos classificadores, e funcionalidades adicionais de análise visual, como redução de dimensionalidade.

Além disso neste trabalho para todas as visualizações criadas utilizando o *VCS* foi

utilizada a técnica *Least Squares Projection* (LSP) (PAULOVICH et al., 2008) disponibilizada no sistema usando 10% do número total de instâncias dos dados como o número de pontos de controle. Também foi utilizada a funcionalidade de coloração dos pontos da projeção, sendo que cada cor aqui representa um evento específico presente nos vídeos utilizados.

4.2 Vídeos

Nesta seção são descritos os vídeos de segurança utilizados na análise da técnica proposta. O primeiro vídeo foi obtido do conjunto de dados CDNet2014 (GOYETTE et al., 2012), um repositório que possui vídeos contendo gravações de ambientes internos e externos para a análise de movimentos no cenário, além de incluir vários cenários de segurança. O segundo vídeo foi obtido do conjunto de dados chamado **VIRAT** (OH et al., 2011), uma coleção de vídeos contendo a interação entre um humano e um veículo. As subseções seguintes descrevem os vídeos.

4.2.1 Sofa

Sofa é um vídeo que possui 2750 quadros distribuídos em 91,7 segundos, com uma taxa de 30 quadros por segundo que mostram pessoas entrando e saindo da cena, objetos sendo deixado sobre um sofá e removidos de cena. A Figura 7 apresenta alguns quadros do vídeo e a Tabela 1 apresenta os eventos presentes do vídeo. A análise desse tipo de vídeo foi realizada pelo fato de os eventos apresentarem múltiplos objetos sendo movimentados, que se relacionam em cenários de segurança a eventos de roubo de objetos, ataques terroristas com bombas, entre outros.



(a) Cena vazia.

(b) Pessoa entra em cena carregando um objeto.

(c) Objetos abandonados.

Figura 7 – Alguns quadros do vídeo **Sofa**.

4.2.2 VIRAT

O vídeo **VIRAT_S_010108_01_000570_000718**, nomeado no trabalho como **VIRAT**, é composto por 3461 quadros distribuídos em 144 segundos, com uma taxa de

Tabela 1 – Tabela de eventos do vídeo **Sofa**

Segmento	Intervalos de quadros	Descrição
1	1-88 e 2747-2750	Cena vazia
2	89-461	Uma pessoa carregando um objeto
3	462-606 e 2143-2283	Um objeto abandonado
4	607-1001	Segundo objeto e pessoa sentada
5	1002-1126	Dois objetos abandonados
6	1127-1434	Terceiro objeto e pessoa sentada
7	1435-1518 e 1729-1878	Três objetos abandonados
8	1519-1728 e 2284-2355 e 2532-2746	Uma pessoa em cena
9	1879-2142	Retirando dois objetos
10	2356-2531	Duas pessoas em cena

30 quadros por segundo, cujo cenário é um estacionamento em que ocorrem vários eventos individuais, como pessoas diferentes caminhando pelo estacionamento, diferentes veículos entrando e saindo de cena e uma pessoa retirando um carro estacionado.

A Figura 8 apresenta quadros de exemplo do evento do carro estacionado saindo do estacionamento e a Tabela 2 apresenta os eventos relacionados a essas ações. Esse vídeo representa eventos em que um objeto é removido de seu local original, tais ações em cenários de segurança estão relacionadas a roubos e assassinatos.

Figura 8 – Evento do carro estacionado saindo do estacionamento do vídeo **VIRAT**.Tabela 2 – Tabela de eventos relacionados a um automóvel saindo do estacionamento do vídeo **VIRAT**

Segmento	Intervalos de quadros	Descrição
1	1-139 e 3189-3461	Estacionamento sem movimento
2	140-730	Moto e ator em cena
3	731-866	Ator chega ao veículo
4	867-1437	Ator entra no veículo
5	1438-2296	Ator dentro do veículo
6	2297-3188	Veículo sai do estacionamento

4.3 Experimentos

Nesta seção serão apresentados os experimentos realizados nos vídeos mencionados.

4.3.1 Avaliação das Técnicas de Extração de características

Esse experimento consiste em avaliar o impacto no *layout* gerado dos vídeos completos ao se utilizar as características extraídas por cada uma das técnicas extratoras de características presentes no projeto proposto, para determinar quais extratores geram *layouts* informativos para análises posteriores. As Figuras 9 e 10 apresentam os *layouts* gerados pelo *VCS* para os vídeos **Sofa** e **VIRAT** completos respectivamente, onde cada ponto representa um quadro do vídeo e cada cor representa um evento presente no vídeo apresentados anteriormente nas Tabelas 1 e 2.

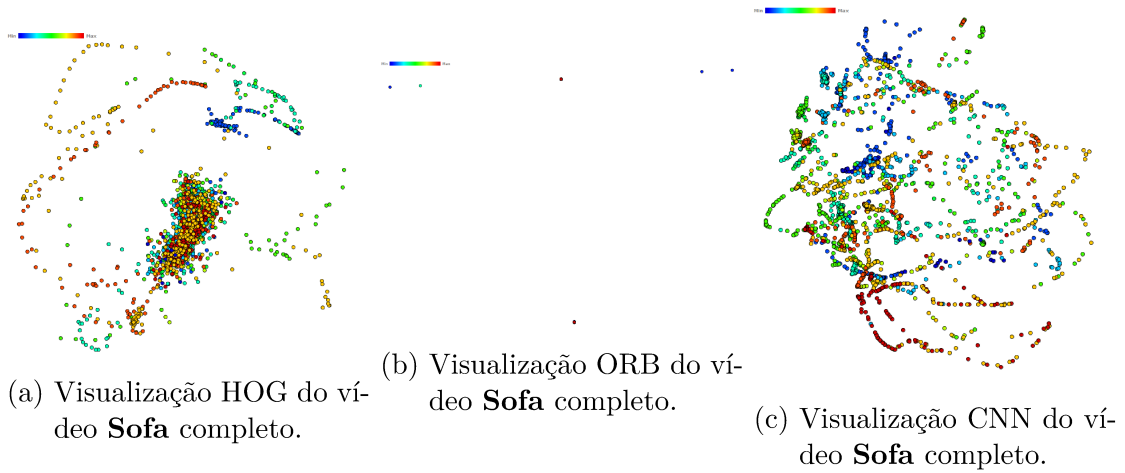


Figura 9 – Visualizações do vídeo **Sofa** completo utilizando as três técnicas extratoras.

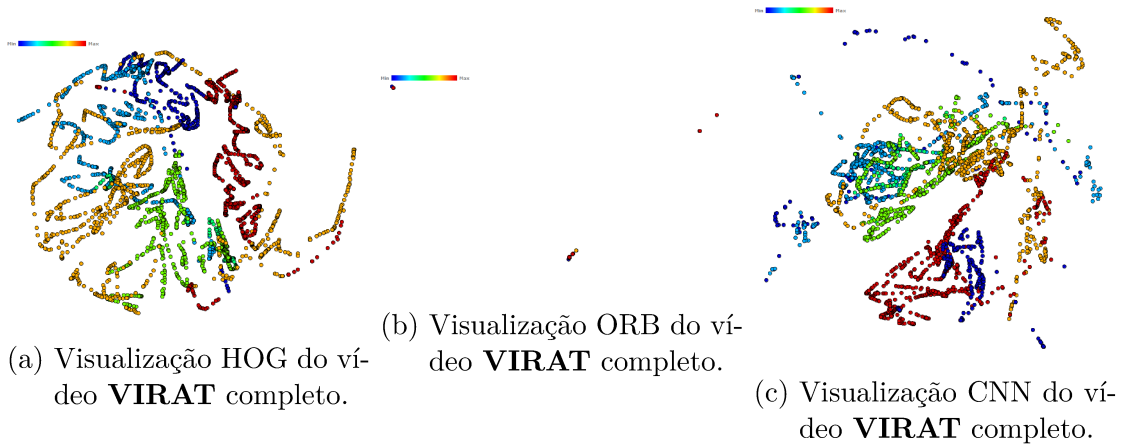


Figura 10 – Visualizações do vídeo **VIRAT** completo utilizando as três técnicas extratoras.

A partir desses *layouts*, é possível notar que os descritores das técnicas HOG e CNN geraram *layouts* com distribuições de pontos que permitem diferenciar cada evento apresentado para o vídeo **VIRAT**, uma vez que os grupos coloridos são distinguíveis. Para o vídeo **Sofa** ambas as técnicas geraram *layouts* em que cada grupo colorido é distinguível, mas há uma sobreposição de pontos e cores no centro para a técnica HOG e uma mistura de cores para a técnica CNN. Tais características estão relacionadas ao

fato de que no vídeo **Sofa** alguns eventos são semelhantes, como os eventos em que objetos são abandonados sobre o sofá que se diferenciam pela quantidade de objetos, bem como quadros de eventos diferentes com movimentações parecidas, como a entrada de uma pessoa carregando um objeto. Por outro lado, os *layouts* gerados pela técnica ORB apresentaram muita sobreposição de quadros, e a percepção de grupos que representem os eventos é muito difícil para ambos os vídeos.

Portanto, para os próximos experimentos, serão utilizadas apenas as técnicas HOG e CNN para avaliação e criação de amostras, dado que a técnica ORB compromete a análise de representatividade visual dos eventos, que é o foco deste trabalho.

4.3.2 Análise Computacional

Neste experimento, procedeu-se à coleta sistemática dos tempos de execução da técnica proposta com o intuito de determinar os impactos dos parâmetros do sistema. Diversas taxas de similaridade foram empregadas, visando aferir os respectivos impactos no tempo de processamento necessário para a geração da amostra final no formato *Points File (.data)*. Os tempos de execução obtidos são mostrados em duas curvas representativas: uma relacionando a Taxa de Similaridade com o Tempo de Execução e a outra relacionando a Percentagem de Seleção da Amostra com o Tempo de Execução.

Para cada vídeo, foram extraídas amostras utilizando taxas de similaridade L de 30%, 50% e 90%, utilizando cada uma das técnicas de extração de características. A Figura 11 mostra os resultados obtidos em relação ao tempo de criação da amostra pela técnica proposta de acordo com a variação da taxa de similaridade.

É possível notar que taxas de similaridade maiores resultam em um tempo de processamento menor. Isso ocorre porque uma taxa de similaridade mais alta faz com que novos agrupamentos sejam formados apenas para quadros com características significativamente diferentes em comparação aos quadros anteriores. Como resultado, menos agrupamentos são criados e, conseqüentemente, menos cálculos de diferença entre o quadro corrente e o centroide dos grupos são realizados. Esse fato é evidente nas curvas de ambos os vídeos referentes as técnicas de extração Rede Neural Convolucional (CNN). Para ambas as técnicas de extração o tempo de processamento foi reduzido de algumas horas, para taxas baixas, para apenas alguns segundos, para taxas de similaridade mais altas.

O mesmo experimento foi executado para analisar o impacto das taxas de porcentagem, que determinam o volume da amostra. Foram utilizadas três porcentagens de seleção diferentes: 30%, 50% e 90%, cujos resultados são mostrados na Figura 12.

É possível notar que o valor da porcentagem de seleção não gera um impacto significativo no tempo de processamento, logo é um parâmetro que se adapta as necessidades

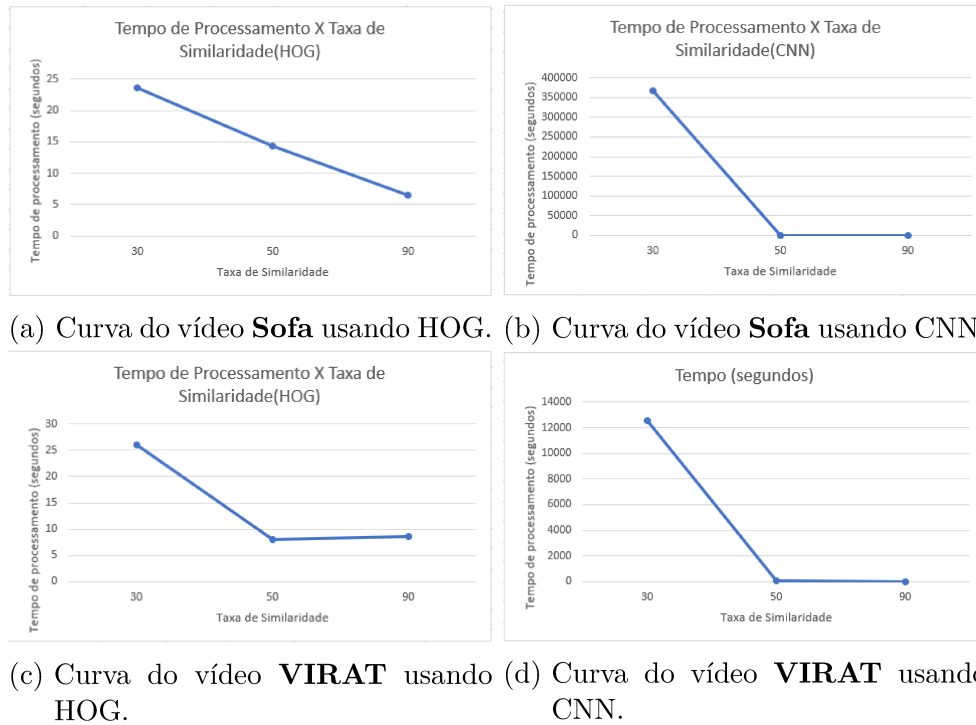


Figura 11 – Curvas Tempo X Taxa de Similaridade da criação das amostras dos vídeos **Sofa** e **VIRAT**.

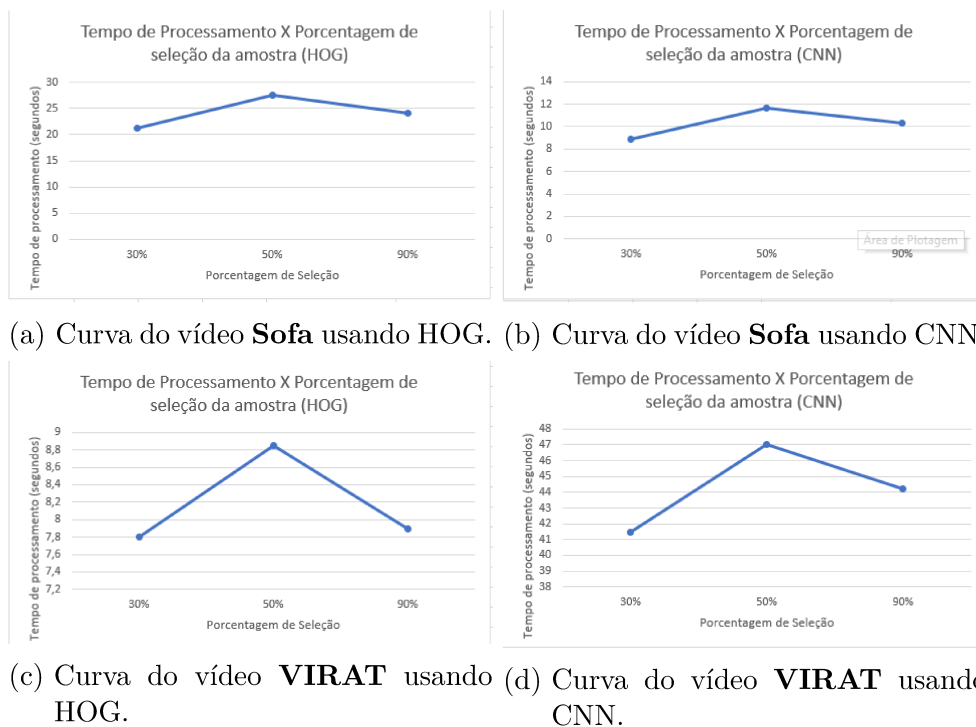


Figura 12 – Curvas Tempo X Porcentagem de seleção da amostra dos vídeos **Sofa** e **VIRAT**.

do usuário independentemente das configurações da máquina que o sistema computacional é executado.

Logo é possível averiguar que ao se aplicar taxas de similaridade adequadas o

tempo de processamento é reduzido consideravelmente, o que é de particular interesse para vídeos de segurança que possuem várias horas de gravação.

4.3.3 Experimento de variação da taxa de similaridade

Neste experimento foi avaliado o impacto da taxa de similaridade na representatividade da amostra gerada pela técnica proposta, a fim de determinar a taxa de similaridade ideal para cada vídeo utilizando ambas as técnicas de extração de características. Para este experimento a porcentagem de quadros da amostra P foi fixado em 60% para se captar nos *layouts* a variação causada pelas diferentes taxas de similaridade.

Para cada vídeo, foram extraídas amostras utilizando taxas de similaridade L de 30%, 50% e 90%, utilizando as técnicas de extração HOG e CNN.

A Figura 13 mostra os *layouts* dos vídeos **Sofa** e **VIRAT** utilizando uma taxa de similaridade de 30%. Para o vídeo **Sofa**, observa-se variações nas amostras resultantes. No caso dos descritores HOG (Figura 13(a)), há uma mudança na forma dos eventos nas extremidades, o que pode ser atribuído à sensibilidade desse tipo de descritor às variações de textura e forma. Já com os descritores CNN (Figura 13(b)), as alterações nos eventos são menos acentuadas, embora ainda presentes, refletindo uma maior estabilidade desse descritor em capturar características essenciais com menos quadros. Em ambos os casos, apesar de algumas distorções, as amostras tendem a manter os padrões gerais dos *layouts* completos do vídeo (Figuras 9(a) e 9(c)). Dessa forma, uma taxa de similaridade de 30% resultou em amostras com reduções de dados e preservem as características dos vídeos analisados.

Quanto ao vídeo **VIRAT**, a organização espacial dos pontos se manteve, embora simplificada pela baixa taxa de similaridade que induz à formação de múltiplos grupos e, conseqüentemente, à seleção de quadros similares para a amostra. A distribuição dos grupos coloridos se preservou parcialmente no *layout* da amostra HOG, como demonstrado na Figura 13(c), visto que algumas cores são pouco representadas no *layout*. Já no *layout* das amostras CNN (Figura 13(d)), apesar da simplificação similar à do HOG, observa-se um predomínio de uma cor nos grupos coloridos. Isso sinaliza que uma taxa de similaridade de 30% resultou numa amostra com super-representação de um evento específico, falhando em retratar adequadamente todos os eventos do vídeo.

A Figura 14 exibe os *layouts* das amostras dos vídeos **Sofa** e **VIRAT** com uma taxa de similaridade de 50%. Para o vídeo **Sofa**, os *layouts* utilizando descritores HOG e CNN, apresentados nas Figuras 14(a) e 14(b), respectivamente, conservam parcialmente os padrões visuais da distribuição de pontos e mantém a distinção entre os agrupamentos de eventos por cor. Para a amostra HOG isso se deu pelo fato de que os há uma simplificação mais acentuada dos pontos dispersos que representam eventos que se repetem ao longo do

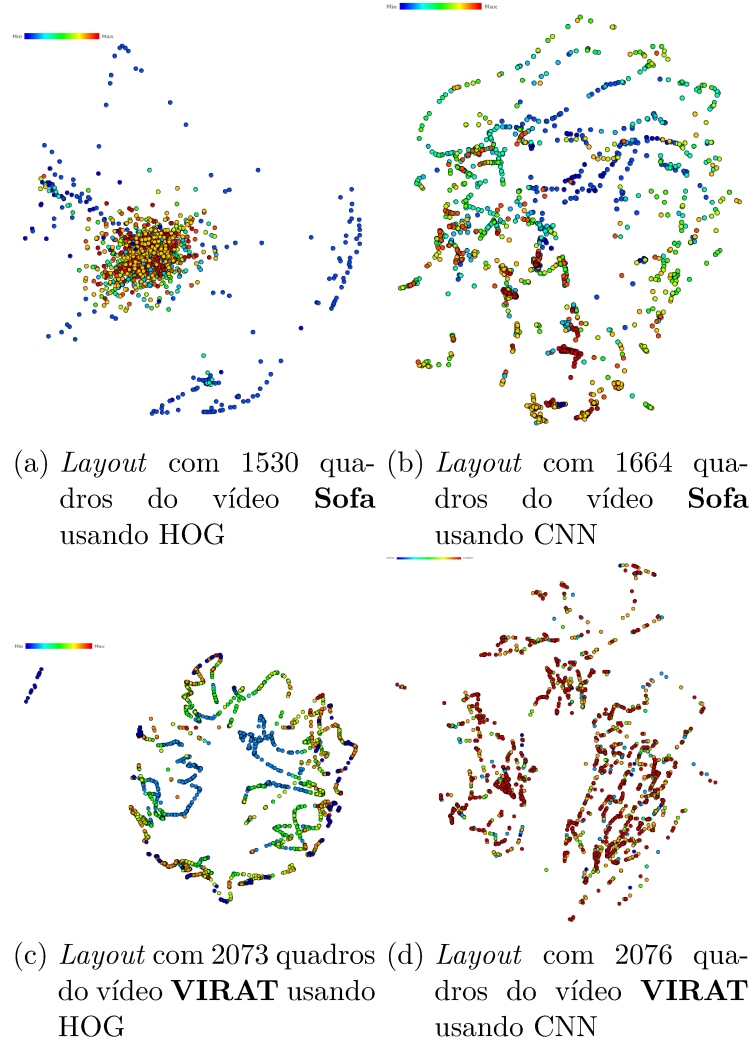


Figura 13 – *Layouts* das amostras dos vídeos **Sofa** e **VIRAT** utilizando uma taxa de similaridade igual a 30%.

vídeo se comparada com o *layout* criado com taxa de similaridade igual a 30%, ou seja a representatividade dos eventos foi prejudicada. Já para CNN há perda de representação de aglomerações centrais dos pontos, visto que para o *layout* utilizando a taxa de similaridade igual a 30% (Figura 13(b)) há uma concentração de 414 quadros na região central enquanto para a taxa de similaridade de 50% a concentração central é reduzida para 226 quadros, ou seja, uma distribuição de pontos simplificada para ambos os descritores.

Quanto ao vídeo **VIRAT**, as Figuras 14(c) e 14(d) mostram que a maior taxa de similaridade favoreceu a correspondência com os *layouts* dos vídeos completos (Figura 10), mantendo a distribuição espacial dos pontos, bem como a distinção dos grupos coloridos que refletem a diversidade das atividades para a amostra utilizando descritores HOG, apesar dos grupos representados pelas cores azul escuro e vermelho estarem mais espalhados, ou seja, possuem uma distribuição mais dispersa em comparação ao *layout* com taxa menor. O *layout* com CNN, em particular, realça a definição dos grupos, indicando que 50% de similaridade otimiza a distinção entre eventos distintos, demonstrado pela presença de

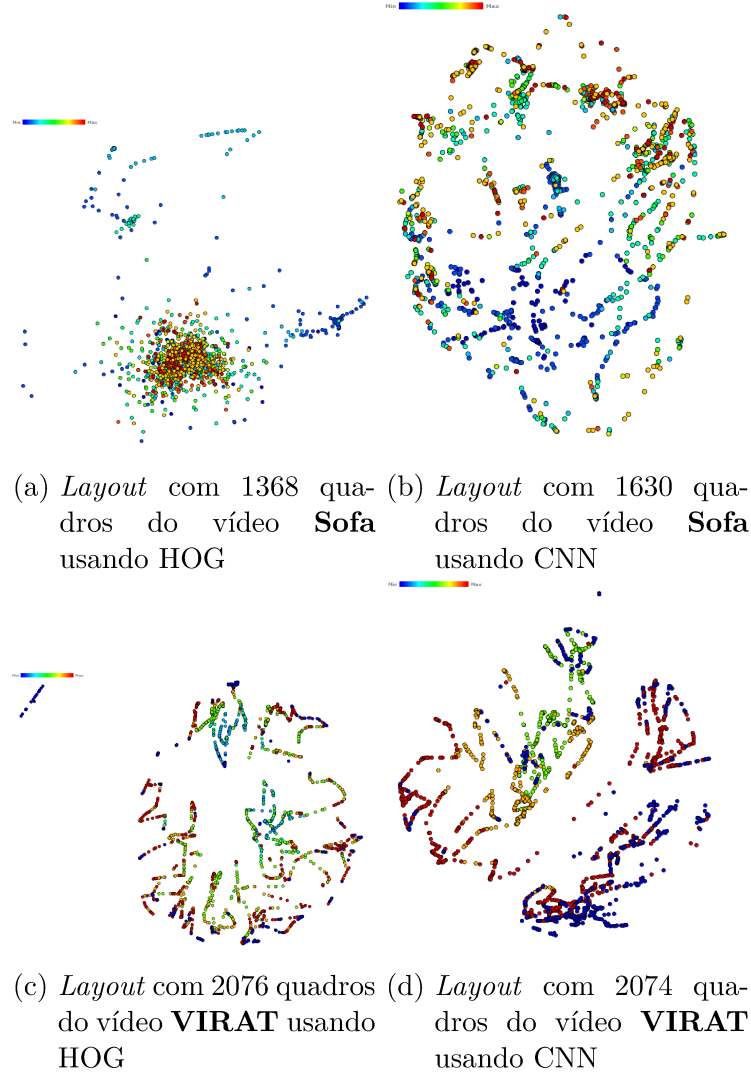


Figura 14 – *Layouts* das amostras dos vídeos **Sofa** e **VIRAT** utilizando uma taxa de similaridade igual a 50%.

grupos de cores distintas ao invés da predominância de uma cor.

A Figura 15 retrata os *layouts* das amostras dos vídeos **Sofa** e **VIRAT** utilizando uma elevada taxa de similaridade de 90%. Comparativamente aos *layouts* dos vídeos completos (Figura 9), notam-se diferenças marcantes.

Para o vídeo **Sofa**, o *layout* com descritores HOG (Figura 14(a)) mostra uma simplificação elevada da distribuição espacial dos pontos com densidade de pontos muito superior à do vídeo completo (Figura 9(a)). Essa simplificação e sobreposição excessiva sugere que detalhes importantes dos eventos do vídeo foram perdidos. Em contraste, a amostra com descritores CNN (Figura 15(b)) apresenta diferenças quanto aos padrões de posições de pontos em relação ao vídeo completo (Figura 9(b)), em virtude da alta taxa de similaridade houve a criação de menos grupos, logo cada grupo passou a englobar quadros com diferenças significativas entre si o que gerou uma distribuição significativamente mais dispersa.

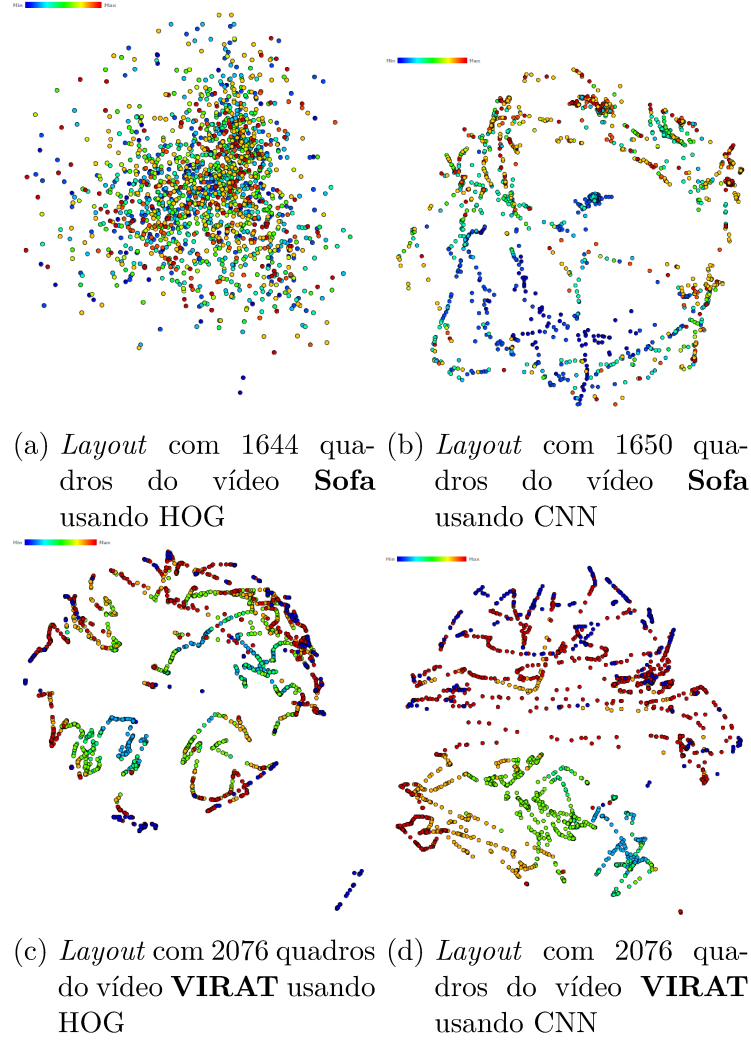


Figura 15 – *Layouts* das amostras dos vídeos **Sofa** e **VIRAT** utilizando uma taxa de similaridade igual a 90%

Quanto ao vídeo **VIRAT**, a amostra HOG (Figura 15(c)) revela um *layout* similar ao gerado para a taxa de similaridade de 50%, mantendo a distribuição espacial dos pontos e grupos coloridos, demonstrando assim uma invariância quanto a variação da taxa de similaridade médias e altas. Para o *layout* da amostra CNN (Figura 15(d)) comparando-se com o *layout* do vídeo completo CNN (Figura 10(b)), a alta taxa de similaridade resultou em uma distribuição espacial e representação de cores com uma similaridade maior ao vídeo original em comparação com *layouts* de taxas menores.

A técnica HOG funcionou bem para extrair características com taxas de similaridade abaixo de 50%, mantendo as características visuais dos vídeos originais, como foi demonstrado no vídeo **Sofa**, que mostra um ambiente interno com movimentações claras. No entanto, taxas de similaridade superiores a cinquenta por cento resultam em uma representação excessivamente simplificada, visto que grande parte dos padrões visuais do *layout* do vídeo completo foram perdidos, assim a distribuição dos pontos de cada evento foi afetada. Da mesma forma, utilizando CNN, observa-se que a perda de detalhes dos

padrões visuais do vídeo original se intensifica com taxas de similaridade acima de 50%, com a perda de formas no *layout* da amostra.

Por outro lado, no vídeo **VIRAT**, que é caracterizado por um ambiente externo com variações dinâmicas, os layouts são observados com HOG com variações entre 50% e 90% em taxas de similaridade. Isso indica que a distribuição de pontos e grupos coloridos é consistente com o layout completo do vídeo mostrado na Figura 10(a). Mas uma taxa de trinta por cento tem um impacto significativo na representação de grupos coloridos, com algumas cores tendo apenas alguns pontos representativos. A CNN mostra a mesma tendência: taxas de similaridade mais altas produzem *layouts* com padrões visuais mais semelhantes aos do vídeo original, demonstrando a capacidade dessa técnica de lidar com a complexidade visual de um ambiente dinâmico.

Com base nesses achados, adotou-se uma taxa de 30% para amostras do vídeo **Sofa** utilizando HOG e CNN. Para o vídeo **VIRAT**, utilizou-se uma taxa de 90% com HOG e 90% com CNN para os próximos experimentos.

4.3.4 Experimento de variação da porcentagem de quadros da amostra

Este experimento examinou como a porcentagem de seleção de quadros da amostra afeta a representatividade das amostras produzidas pelo método sugerido, com o objetivo de averiguar como a porcentagem afeta a formatação dos *layouts* das amostras. Embora a formação dos grupos ou a seleção dos eventos para a amostra não sejam alterados pela porcentagem de quadros, ela tem um efeito direto na quantidade de pontos presentes, como resultado, afetando as formas e padrões dos grupos no *layout*. Para o experimento, a taxa de similaridade foi fixada em 50%. Isso permitiu identificar as mudanças causadas pela variação na porcentagem de seleção nos *layouts*.

Para cada vídeo, foram extraídas amostras utilizando taxas de seleção P de 30%, 50% e 90%, para as técnicas de extração HOG e CNN.

A Figura 16 apresenta os *layouts* referentes as amostras de ambos os vídeos para as três taxas de seleção utilizando a técnica de extração HOG. Ao analisar a quantidade de quadros e a configuração dos pontos em cada *layout*, observa-se que, para ambos os vídeos, o incremento na porcentagem de quadros da amostra P eleva proporcionalmente o número de quadros na amostra final. Isso fica evidente visto nos *layouts* que representam amostras com uma porcentagem de quadros de 90%, como ilustrado pela Figura 16(c) para o vídeo **Sofa** e pela Figura 16(f) para o vídeo **VIRAT**. Nestes também é possível notar que para a porcentagem de quadros de 30% há uma simplificação exagerada das formas presentes nos *layouts* em comparação com os *layouts* dos vídeos completos (Figuras 9 e 10), em que há perda de muitos detalhes da formatação visual do *layout*.

A Figura 17 apresenta os *layouts* resultantes das amostras dos vídeos **Sofa** e **VI-**

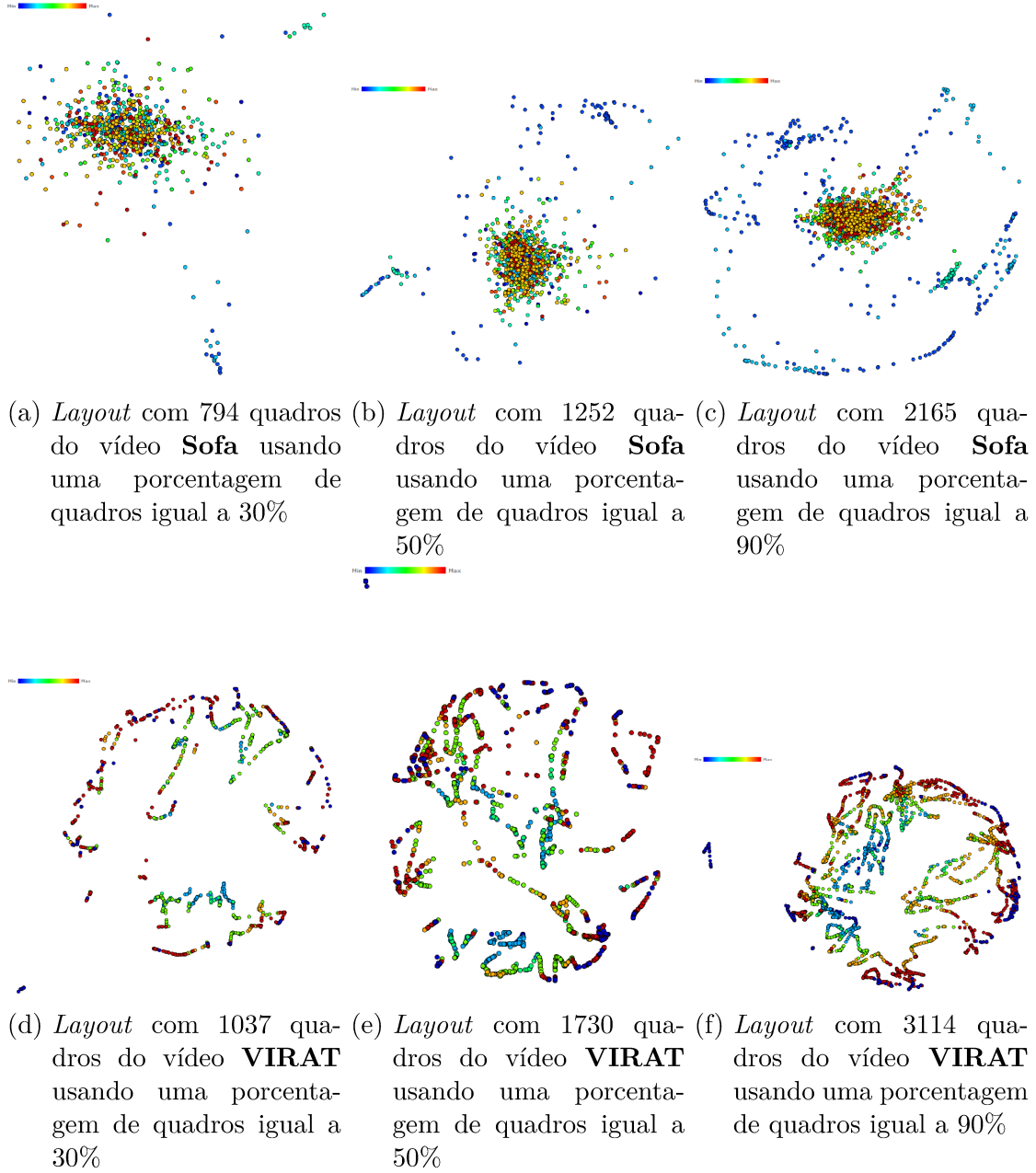


Figura 16 – *Layouts* das amostras dos vídeos **Sofa** e **VIRAT** para variação da porcentagem de quadros da amostra utilizando HOG

RAT, aplicando-se três taxas de seleção distintas com o uso da técnica de extração CNN. A análise da distribuição dos pontos nos *layouts* é possível notar que, assim como observado para a técnica HOG, a porcentagem de quadros igual a 30% (Figuras 17(a) e 17(b)) gera *layouts* com formas e distribuição de grupos simplificados, ou seja, há perda da complexidade dos eventos em sua representação.

Nos *layouts* de amostras com 90% de seleção, ilustrados pelas Figuras 16(c) para o vídeo **Sofa** e 17(f) para o vídeo **VIRAT**, mostram uma representação mais forte dos grupos e sua forma e exibição. Isso denota uma seleção mais ampla da distribuição dos grupos e seus eventos quando comparado às amostras originadas de taxas de seleção

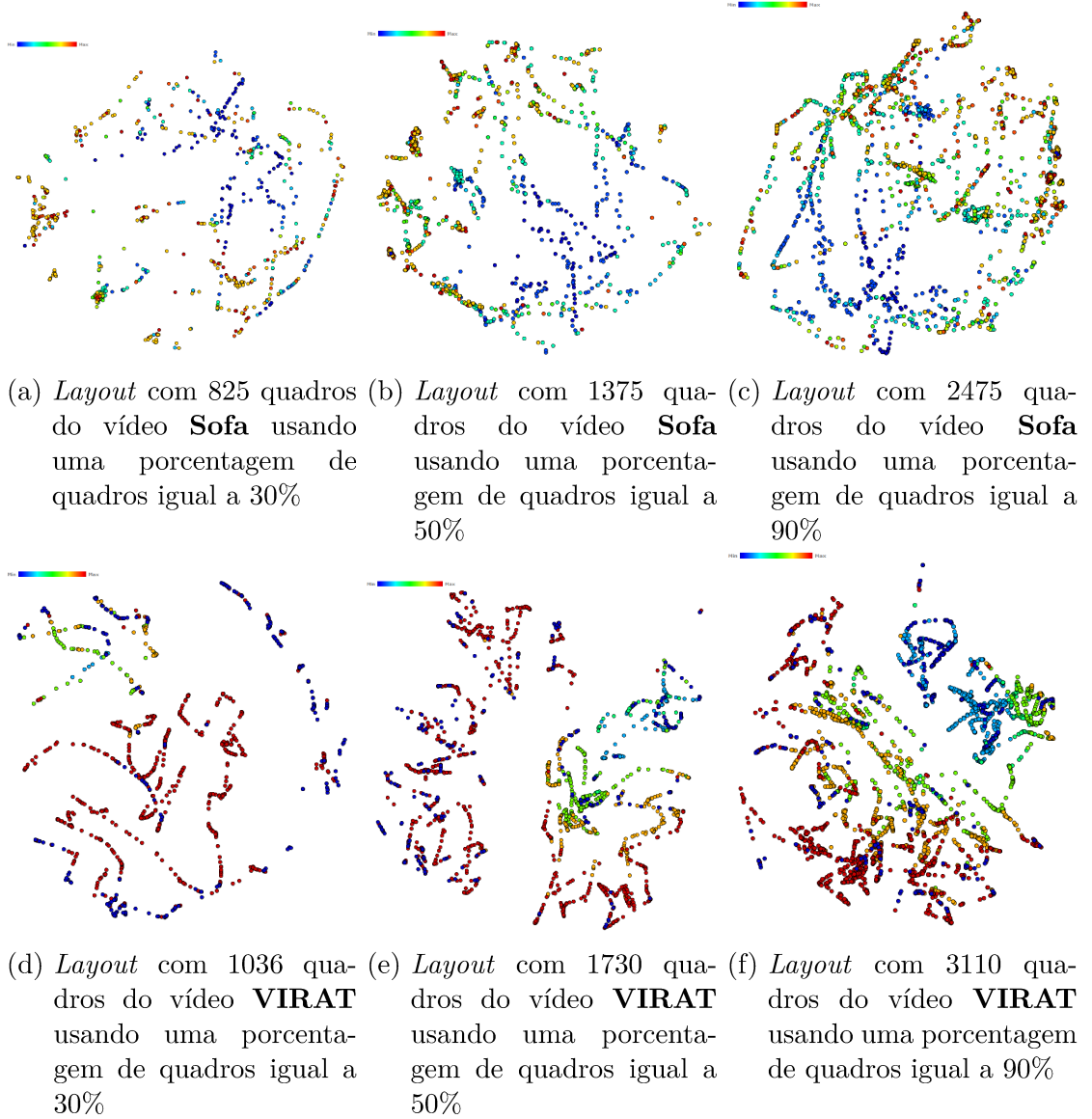


Figura 17 – *Layouts* das amostras dos vídeos **Sofa** e **VIRAT** para variação da porcentagem de quadros da amostra utilizando CNN

menores, evidenciado pela comparação com os *layouts* correspondentes a taxas de 30% e 50%, mostrados nas Figuras 17(a), 17(b), 17(d) e 17(e).

Logo a variação na porcentagem de quadros das amostras tem um impacto significativo sobre as características dos *layouts* gerados pelos vídeos **Sofa** e **VIRAT**. Aumentar a porcentagem de quadros contribui para uma maior representação dos grupos, o que é refletido pela maior densidade de pontos nos *layouts* com taxas mais altas. Esse fenômeno é consistente em ambientes internos e externos, representados pelos vídeos analisados, e demonstra a influência direta da porcentagem de quadros na construção das amostras.

Os resultados mostram que uma porcentagem de quadros mais alta gera amostras com *layouts* que possuem a disposição, forma e formatação dos grupos coloridos mais semelhantes às do vídeo original, enquanto uma porcentagem de quadros mais baixa resulta

em *layouts* simplificados que afetam aspectos como a distribuição dos quadros pertencentes a um evento recorrente durante o vídeo. Contudo, deve-se ponderar a necessidade de compactação de dados versus a preservação dos detalhes, visto que a porcentagem de quadros não afeta a formação dos grupos de quadros semelhantes da técnica proposta e por consequência não afeta a seleção de eventos (grupos coloridos do *layout*). Taxas de seleção próximas a 50% emergem como um ponto de equilíbrio interessante, proporcionando uma redução substancial na quantidade de dados, enquanto mantém a fidelidade na distribuição espacial dos pontos, bem como a representação dos eventos dos vídeos dados pelas diferentes cores, ao se comparar com os *layouts* dos vídeos completos presentes nas Figuras 9 e 10.

Assim para o próximo experimento foi mantida a porcentagem de quadros $P = 60\%$ usada em experimentos anteriores.

4.3.5 Análise da Representatividade Amostral

Neste estudo, é avaliada a representatividade dos eventos em *layouts* obtidos por meio da técnica proposta. A análise foi realizada visualmente, observando a coloração dos pontos nos *layouts*. Além disso, os *layouts* das amostras geradas pela técnica proposta foram comparados com *layouts* construídos a partir de duas técnicas de amostragem distintas: a primeira consiste em selecionar quadros de maneira aleatória ao longo do vídeo, e a segunda selecionando uma porcentagem específica de quadros de cada segundo do vídeo. Para a amostragem aleatória foram utilizados 60% dos quadros dos vídeos originais e para a amostragem por segundo foram selecionados 18 quadros para cada segundo dos vídeos (60%). Para a avaliação da técnica, foi empregada uma porcentagem de quadros igual a 60%, além das taxas de similaridade determinadas na seção 4.3.3.

A Figura 18 apresenta os *layouts* dos vídeos **Sofa** e **VIRAT** completos para descritores HOG e CNN e a Figura 18 apresenta os *layouts* das amostras de ambos os vídeos gerados pela técnica proposta para os parâmetros mencionados anteriormente.

Tanto para o vídeo **Sofa**, quanto o **VIRAT** as amostras geradas utilizando ambas as técnicas de extração de características HOG e CNN apresentaram uma representatividade eficaz dos eventos dos vídeos como apresentados nos *layouts* na Figura 19. Todas as amostras capturaram todos os eventos de ambos os vídeos apresentados nas Tabelas 1 e 2, uma vez que todos os grupos coloridos estão presentes nos *layouts* dos vídeos completos (Figura 18), com uma quantidade de quadros reduzidas. Apesar das amostras representarem efetivamente os eventos de ambos os vídeos, os *layouts* de cada uma apresentam diferenças em relação aos *layouts* dos vídeos completos.

Observa-se que, para o vídeo **Sofa** utilizando a técnica HOG, houve uma maior compactação dos grupos coloridos, devido a menor representação de pontos dispersos

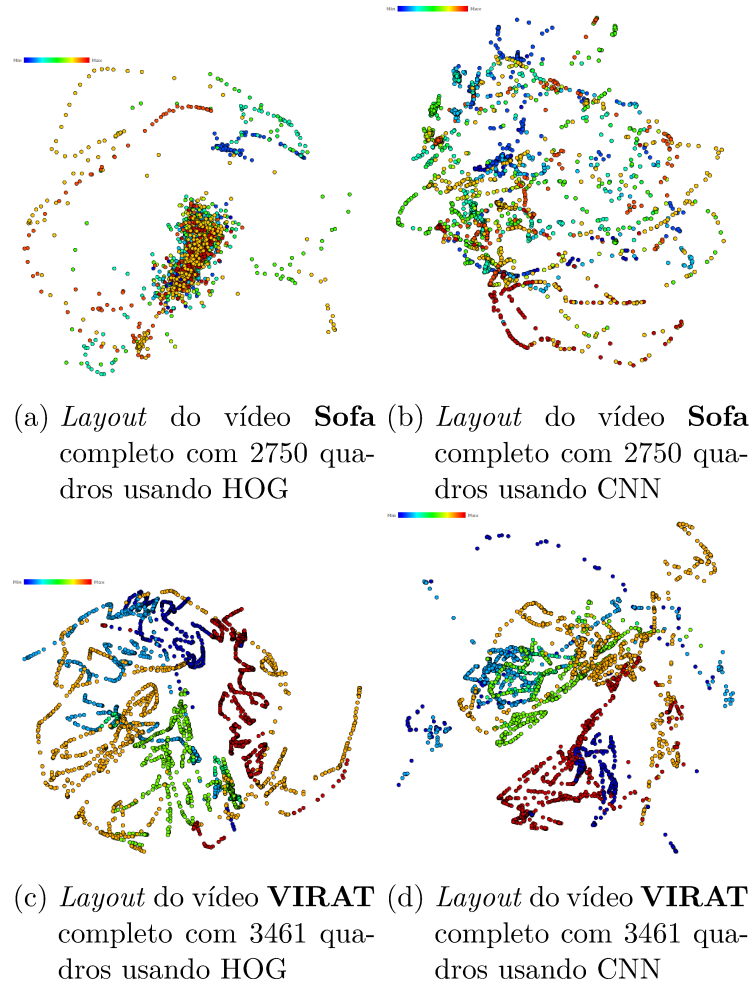


Figura 18 – *Layouts* dos vídeos **Sofa** e **VIRAT** completos com HOG e CNN

nas fronteiras do *layout*, apesar disso o *layout* da amostra representa efetivamente os 10 eventos presentes no vídeo. Já com o uso da técnica CNN, os eventos, embora todos representados, demonstraram uma tendência a uma dispersão de pontos ligeiramente maior comparado ao vídeo completo, mas ainda preserva a distinção geral entre eles.

O vídeo **VIRAT** utilizando HOG apresenta assim como a amostra usando CNN do vídeo **Sofa** uma dispersão de pontos, porém mais acentuada, visto que para cada cor ao invés de seus pontos estarem aglomerados em uma posição os mesmos estão dispersos na fronteira do *layout* da amostra. Apesar disso, todos os 6 eventos são representados de forma distinguível. Já o *layout* CNN apresenta uma ligeira sobreposição de alguns pontos de eventos adjacentes, o que não compromete a identificação deles.

Assim, para vídeos que apresentam cenários internos, com movimentações de pessoas e objetos bem definidos e pouca variação na iluminação, como o vídeo **Sofa**, a técnica de amostragem proposta é capaz de gerar amostras representativas de todos os eventos do vídeo ao adotar uma taxa de similaridade (L) com valor próximo a 30%. Já para vídeos que exibem cenários externos, com movimentações variáveis e alterações de iluminação, como o vídeo **VIRAT**, a técnica de amostragem proposta gera amostras representativas

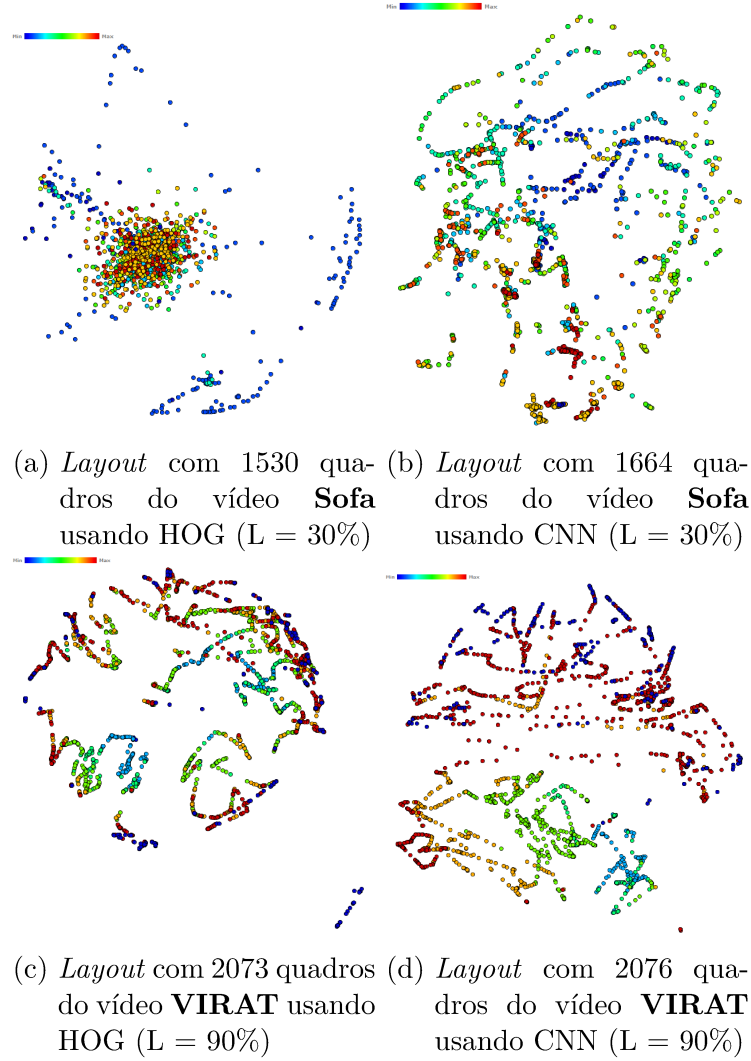


Figura 19 – *Layouts* das amostras da técnica proposta dos vídeos **Sofa** e **VIRAT** para HOG e CNN

de todos os eventos do vídeo com taxas de similaridade elevadas, aproximando-se de 90%. Logo em cenários de segurança a técnica proposta mostrou-se eficiente na criação de amostras que representem ações variadas, como a interação entre pessoa e objeto, movimento de pessoas, movimentos de veículos e alteração na posição de objetos de modo a acelerar a análise posterior desses eventos.

Em relação a comparação dos *layouts* criados pela técnica proposta com a técnicas de amostragem aleatória e por segundo tem-se a Figura 20 que apresenta os *layouts* para ambos os vídeos referentes às amostras criadas a partir das técnicas de amostragem aleatória e amostragem por segundo.

Ao comparar os *layouts* criados pela técnica proposta com os gerados através da amostragem aleatória e por segundo, constata-se que a abordagem proposta neste trabalho oferece resultados promissores. Para o vídeo **Sofa**, os *layouts* gerados pela técnica proposta e apresentados na Figura 19 exibem uma distribuição uniforme de eventos, representados

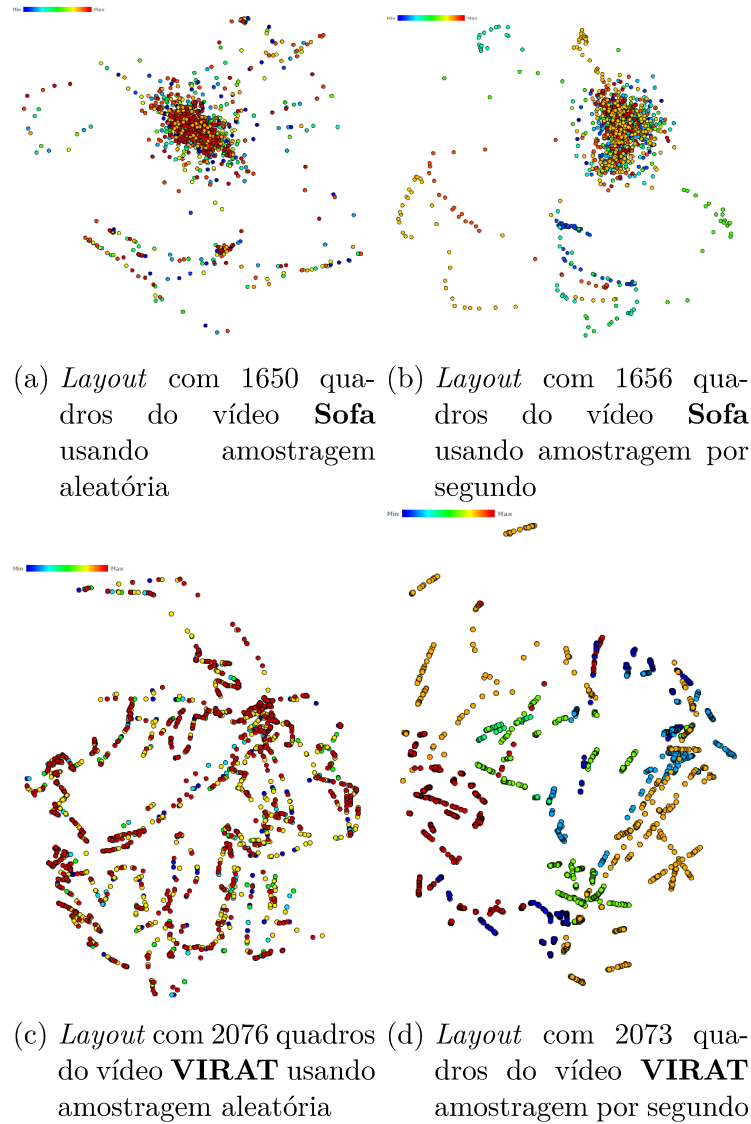


Figura 20 – *Layouts* das amostras dos vídeos **Sofa** e **VIRAT** utilizando amostragem aleatória e amostragem por segundo

por cores distintas, refletindo um equilíbrio na representatividade dos eventos ao longo do vídeo. Essa uniformidade contrasta com os *layouts* resultantes da amostragem aleatória Figura 20(a), onde uma predominância de algumas cores indica uma super-representação dos eventos de maior duração, introduzindo um possível viés na interpretação dos dados.

Adicionalmente, ao observar os *layouts* do vídeo **VIRAT**, é possível notar novamente a capacidade da técnica proposta de evitar o viés de super-seleção, em comparação com a amostragem aleatória que tende a sobrevalorizar determinados eventos (20(c)). É importante pontuar que a seleção aleatória de quadros pode levar a uma percepção desequilibrada da importância dos eventos, visto que eventos longos e contínuos possuem uma maior probabilidade de ter vários de seus quadros selecionados comprometendo a análise qualitativa do conteúdo do vídeo.

Por fim, os *layouts* originados da técnica de amostragem por segundo para am-

bos os vídeos (Figura 20(b) e 20(d)) capturam uma visão mais distribuída dos eventos ao longo do tempo. Entretanto, a técnica de amostragem desenvolvida e utilizada neste estudo se destaca por fornecer um panorama equilibrado e abrangente dos eventos, independentemente de sua duração. A capacidade de ajustar a taxa de similaridade permite refinar a representatividade dos eventos, garantindo que tanto os de curta quanto os de longa duração sejam representados nos *layouts* de forma proporcional. Assim, a técnica proposta demonstra ser superior às metodologias convencionais de amostragem de quadros de vídeos, assegurando uma visão mais holística e precisa dos eventos contidos nos vídeos analisados.

5 Conclusão

Este trabalho apresentou uma técnica de amostragem de quadros para vídeos de vigilância, empregando a formação de grupos com base na similaridade e escolhendo os quadros mais representativos de cada grupo para compor a amostra.

Foram utilizados dois descritores de características para descrever os quadros dos vídeos, o descritor HOG e o CNN. A técnica proposta possui uma flexibilidade considerável para gerar amostras representativas que facilitem e acelerem análises visuais dos eventos presentes em vídeos de segurança completos através de ferramentas visuais.

A eficácia foi demonstrada através da comparação dos *layouts* gerados pelo *VCS* para as amostras criadas pela abordagem proposta e dos *layouts* das técnicas de amostragem aleatória e por segundo. A abordagem proposta consegue capturar a essência dos eventos de vídeo com uma distribuição mais homogênea e completa dos eventos, evitando a super-representação de eventos prolongados típicos da amostragem aleatória e proporcionando uma representação temporal mais fiel do que a amostragem por segundo.

O trabalho contribui para o campo de processamento de vídeos e análise visual de eventos de vídeos ao propor uma técnica eficiente de amostragem de quadros, que pode ser aplicada para reduzir o volume de dados a serem analisados e visualizados sem comprometer a representatividade dos eventos importantes. Para pesquisas futuras, sugere-se a exploração de outras técnicas de extração de características e a aplicação da abordagem proposta em diferentes contextos de vídeos de segurança, a fim de validar sua eficácia e adaptabilidade.

Referências

- ACHARYA, A. S.; PRAKASH, A.; SAXENA, P.; NIGAM, A. Sampling: Why and how of it. **Indian Journal of Medical Specialties**, v. 4, n. 2, p. 330–333, 2013. Disponível em: <<https://doi.org/10.7713/ijms.2013.0032>>. Citado na página 10.
- AUDET, S.; CONTRIBUTORS. **JavaCV: Java interface to OpenCV, FFmpeg, and more**. 2024. Acesso em: 20 de maio de 2024. Disponível em: <<https://github.com/bytedeco/javacv>>. Citado na página 20.
- BOLFARINE, H.; BUSSAB, W. de O. **Elementos de amostragem**. [S.l.]: Editora Blucher, 2005. Citado na página 10.
- BURGOS-ARTIZZU, X. P.; DOLLÁR, P.; LIN, D.; ANDERSON, D. J.; PERONA, P. Social behavior recognition in continuous video. In: IEEE. **2012 IEEE conference on computer vision and pattern recognition**. [S.l.], 2012. p. 1322–1329. Citado na página 15.
- CALONDER, M.; LEPETIT, V.; STRECHA, C.; FUA, P. Brief: Binary robust independent elementary features. In: SPRINGER. **European conference on computer vision**. 2010. p. 778–792. Disponível em: <https://doi.org/10.1007/978-3-642-15561-1_56>. Citado na página 18.
- CAMASTRA, F.; VINCIARELLI, A. **Machine learning for audio, image and video analysis: theory and applications**. Springer, 2015. Disponível em: <<https://doi.org/10.1007/978-1-4471-6735-8>>. Citado na página 13.
- CARREIRA, J.; ZISSERMAN, A. Quo vadis, action recognition? a new model and the kinetics dataset. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [s.n.], 2017. Disponível em: <<https://doi.org/10.1109/CVPR.2017.502>>. Citado na página 13.
- DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: IEEE. **2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)**. [S.l.], 2005. v. 1, p. 886–893. Citado 2 vezes nas páginas 15 e 18.
- DAS, S.; SARKER, A.; MAHMUD, T. Violence detection from videos using hog features. In: IEEE. **2019 4th International Conference on Electrical Information and Communication Technology (EICT)**. 2019. p. 1–5. Disponível em: <<https://doi.org/10.1109/EICT48899.2019.9068754>>. Citado na página 15.
- Eclipse Foundation. **Deeplearning4j: Open-source, Distributed Deep Learning for the JVM**. 2024. Accessed: 2024-04-14. Disponível em: <<https://deeplearning4j.konduit.ai>>. Citado na página 20.
- ELER, D. M.; NAKAZAKI, M. Y.; PAULOVICH, F. V.; SANTOS, D. P.; OLIVEIRA, M. C. F.; NETO, J. E. B.; MINGHIM, R. Multidimensional visualization to support analysis of image collections. In: IEEE. **2008 XXI Brazilian Symposium on**

Computer Graphics and Image Processing. 2008. p. 289–296. Disponível em: <<https://doi.org/10.1109/SIBGRAPI.2008.30>>. Citado na página 21.

ELER, D. M.; NAKAZAKI, M. Y.; PAULOVICH, F. V.; SANTOS, D. P.; ANDERY, G. F.; OLIVEIRA, M. C. F.; NETO, J. B.; MINGHIM, R. Visual analysis of image collections. **The Visual Computer**, Springer, v. 25, p. 923–937, 2009. Disponível em: <<https://doi.org/10.1007/s00371-009-0368-7>>. Citado na página 22.

ESPADOTO, M.; HIRATA, N. S. T.; TELEA, A. C. Deep learning multidimensional projections. **Information Visualization**, SAGE Publications Sage UK: London, England, v. 19, n. 3, p. 247–269, 2020. Disponível em: <<https://doi.org/10.1177/1473871620909485>>. Citado na página 12.

FAN, H.; XU, Z.; ZHU, L.; YAN, C.; GE, J.; YANG, Y. Watching a small portion could be as good as watching all: Towards efficient video classification. In: **IJCAI International Joint Conference on Artificial Intelligence**. [s.n.], 2018. Disponível em: <<https://doi.org/10.24963/ijcai.2018/98>>. Citado na página 14.

FEICHTENHOFER, C. X3d: Expanding architectures for efficient video recognition. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [s.n.], 2020. Disponível em: <<https://doi.org/10.1109/CVPR42600.2020.00028>>. Citado na página 13.

GOYETTE, N.; JODOIN, P.-M.; PORIKLI, F.; KONRAD, J.; ISHWAR, P. Changedetection. net: A new change detection benchmark dataset. In: **IEEE. 2012 IEEE computer society conference on computer vision and pattern recognition workshops**. 2012. p. 1–8. Disponível em: <<https://doi.org/10.1109/CVPRW.2012.6238919>>. Citado na página 23.

International Organization for Standardization. **ISO/IEC 14496-14:2020 Information technology – Coding of audio-visual objects – Part 14: MP4 file format**. 2020. Accessed: 2024-04-14. Disponível em: <<https://www.iso.org/standard/74428.html>>. Citado na página 21.

KABANI, A.; EL-SAKKA, M. R. Object detection and localization using deep convolutional networks with softmax activation and multi-class log loss. In: **SPRINGER. International Conference on Image Analysis and Recognition**. 2016. p. 358–366. Disponível em: <https://doi.org/10.1007/978-3-319-41501-7_41>. Citado 2 vezes nas páginas 3 e 13.

LI, H.; LI, Y.; PORIKLI, F. et al. Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking. In: **BMVC**. [s.n.], 2014. v. 1, n. 2, p. 3. Disponível em: <<https://doi.org/10.5244/C.28.56>>. Citado 3 vezes nas páginas 3, 13 e 14.

NEVES, T. T. d. A. T. **Projeções multidimensionais para a análise de fluxos de dados**. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado na página 12.

OH, S.; HOOGS, A.; PERERA, A.; CUNTOOR, N.; CHEN, C.-C.; LEE, J. T.; MUKHERJEE, S.; AGGARWAL, J.; LEE, H.; DAVIS, L. et al. A large-scale benchmark dataset for event recognition in surveillance video. In: **IEEE. CVPR 2011**. 2011. p. 3153–3160. Disponível em: <<https://doi.org/10.1109/CVPR.2011.5995586>>. Citado na página 23.

- Open Source Vision Foundation. **OpenCV - Open Source Computer Vision Library**. 2024. Accessed: 2024-04-14. Disponível em: <<https://opencv.org>>. Citado na página 20.
- PAIVA, J. G. S.; SCHWARTZ, W. R.; PEDRINI, H.; MINGHIM, R. An approach to supporting incremental visual data classification. **IEEE transactions on visualization and computer graphics**, IEEE, v. 21, n. 1, p. 4–17, 2014. Disponível em: <<https://doi.org/10.1109/TVCG.2014.2331979>>. Citado 2 vezes nas páginas 11 e 22.
- PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 14, n. 3, p. 564–575, 2008. Disponível em: <<https://doi.org/10.1109/TVCG.2007.70443>>. Citado na página 23.
- ROSTEN, E.; DRUMMOND, T. Machine learning for high-speed corner detection. **European conference on computer vision**, Springer, p. 430–443, 2006. Disponível em: <https://doi.org/10.1007/11744023_34>. Citado na página 18.
- RUBLEE, E.; RABAU, V.; KONOLIGE, K.; BRADSKI, G. Orb: An efficient alternative to sift or surf. In: IEEE. **2011 International conference on computer vision**. 2011. p. 2564–2571. Disponível em: <<https://doi.org/10.1109/ICCV.2011.6126544>>. Citado na página 18.
- SADIKU, M.; SHADARE, A. E.; MUSA, S. M.; AKUJUOBI, C. M.; PERRY, R. Data visualization. **International Journal of Engineering Research And Advanced Technology (IJERAT)**, v. 2, n. 12, p. 11–16, 2016. Citado na página 11.
- SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: **3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings**. [S.l.: s.n.], 2015. Citado na página 18.
- SONG, J.; GAO, S.; ZHU, Y.; MA, C. A survey of remote sensing image classification based on cnns. **Big earth data**, Taylor & Francis, v. 3, n. 3, p. 232–254, 2019. Disponível em: <<https://doi.org/10.1080/20964471.2019.1657720>>. Citado na página 13.
- TRAN, D.; BOURDEV, L.; FERGUS, R.; TORRESANI, L.; PALURI, M. Learning spatiotemporal features with 3d convolutional networks. In: **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**. [s.n.], 2015. Disponível em: <<https://doi.org/10.1109/ICCV.2015.510>>. Citado na página 13.
- WANG, L.; XIONG, Y.; WANG, Z.; QIAO, Y.; LIN, D.; TANG, X.; GOOL, L. V. Temporal segment networks: Towards good practices for deep action recognition. In: SPRINGER. **European conference on computer vision**. 2016. p. 20–36. Disponível em: <https://doi.org/10.1007/978-3-319-46484-8_2>. Citado na página 13.
- WU, W.; HE, D.; TAN, X.; CHEN, S.; WEN, S. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In: _____. **2019 IEEE/CVF International Conference on Computer Vision (ICCV)**. IEEE, 2019. p. 6222–6231. Disponível em: <<https://doi.org/10.1109/ICCV.2019.00632>>. Citado 3 vezes nas páginas 3, 14 e 15.

ZHI, Y.; TONG, Z.; WANG, L.; WU, G. Mgsampler: An explainable sampling strategy for video action recognition. In: _____. **2021 IEEE/CVF International Conference on Computer Vision (ICCV)**. IEEE, 2021. p. 1513–1522. Disponível em: <<https://doi.org/10.1109/ICCV48922.2021.00154>>. Citado na página 15.