Fusão de Dados de Vídeos RGB e Pontos-Chaves para Classificação de Movimentos Gerais de Bebês

Matheus Palheta[1], Giovanna Santos[1], Ayrles Mendonça[2], Paulo Gonçalves[1], Rafael Albuquerque[1], Eduardo Souto[1], Eulanda M. dos Santos[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)

[2]Faculdade de Educação Física e Fisioterapia – UFAM
Av. Gen. Rodrigo Octávio 6200, Coroado I – 69080-900 – Manaus – AM – Brasil

{matheus.palheta, giovanna.andrade, paulo, esouto}@icomp.ufam.edu.br,

{rafael.albuquerque,emsantos}@icomp.ufam.edu.br, ayrles@ufam.edu.br

Abstract. *The detection of atypical spontaneous movements of babies, called general movements, can help in the early identification of neurodevelopmental disorders and consequent early intervention. General movements can be evaluated using video-based data by machine learning techniques. In this paper, babies' movements are classified into typical and atypical using a video dataset collected for this work. Two data channels are generated from each video: the RGB videos and keypoints. The data provided by each channel is used to train two 3D convolutional neural network classification models, one for each channel. The two models are combined using a fusion function. The results show that the fusion of the models achieves higher classification rates when compared to the rates obtained by the models trained with data provided by each channel individually.*

Resumo. *A detecção de atipicidade na execução dos movimentos espontâneos de bebês, chamados de movimentos gerais, pode ajudar na identificação precoce de distúrbios do neurodesenvolvimento e consequente início precoce de intervenção. Os movimentos gerais podem ser avaliados por meio de vídeos utilizando técnicas de aprendizado de máquina. Neste trabalho, nós classificamos os movimentos dos bebês em típicos e atípicos a partir de uma base de vídeos própria, dos quais são gerados dois canais de dados: os próprios vídeos RGB e pontos-chaves. Os dados providos por cada canal são empregados para treinar dois modelos de classificação baseados em rede neural de convolução 3D, um para cada canal. Os modelos são combinados via função de fusão. Os resultados mostram que a fusão dos modelos alcança taxas de classificação superiores às taxas obtidas pelos modelos treinados com dados providos por cada canal individualmente.*

1. Introdução

Os comprometimentos do neurodesenvolvimento são alterações que envolvem o crescimento e desenvolvimento do sistema nervoso central, levando a repercussões na funcionalidade dos indivíduos. A *Qualitative Assessment of General Movements* de Prechtl (GMA) é um método de avaliação de bebês, que se baseia na observação e avaliação qualitativas da movimentação geral e espontânea. A observação dos movimentos espontâneos

pode determinar presença de alterações neuromotoras e comprometimentos de neurodesenvolvimento relacionados a integridade do sistema nervoso central (SNC), pois a qualidade dos movimentos é modulada por vias corticoespinais ou reticuloespinais e pode ser afetada por alterações dessas estruturas. Usualmente, os movimentos gerais ou "*general movements*" (GMs) podem ser caracterizados em dois tipos, de acordo com a idade do lactente: os *Writhing Movements* (WVs), de 28 semanas até 2 meses; e os *Fidgety Movements* (FMs), de 3 a 5 meses. A fase entre 2 e 3 meses é transicional e de difícil análise, pois engloba componentes de WMs e FMs, além da aquisição de movimentação voluntária cortical [Einspieler et al. 2016].

Tradicionalmente os GMs possuem repertório rico e complexo e uma organização espaço-temporal específica. Sua análise compreende a observação do movimento de todo o corpo, a manifestação da sequência e variação dos movimentos de pescoço, braço, tronco e pernas, assim como a avaliação da velocidade, variabilidade, fluência, intensidade, complexidade e amplitude desses movimentos [Aizawa et al. 2021].

Nesse sentido, a atipicidade na execução dos GMs pode indicar distúrbios do neurodesenvolvimento, envolvendo a predição de risco de paralisia cerebral ou até mesmo a presença de deficiências cognitivas [Aizawa et al. 2021]. Por essa razão, a GMA é considerada uma ferramenta de avaliação recomendada para acompanhamento de bebês, especialmente prematuros ou com fatores de riscos atrelados, como: baixo peso, exposição à infecções intrautero, associação com idade materna avançada, pré-eclampsia, oligodrâmnia, entre outros [Raghuram et al. 2022].

Por meio da GMA, através da análise de vídeo por profissional capacitado e experiente na avaliação infantil, é possível realizar uma predição de risco de atipicidades, ainda numa fase precoce e de alta neuroplasticidade para os bebês monitorados, o que possibilita a intervenção oportuna, conhecida também como precoce, e amplia as chances de evolução e minoração de comprometimentos [Leo et al. 2022]. Assim, quanto mais precoce a identificação, maiores as chances de que estimulações específicas produzam modificações positivas no SNC. Nessa perspectiva, compreender os GMs, em especial os WMs, é imperativo e deve ser incentivado como prática cotidiana de avaliação infantil. Entretanto, a aplicação da GMA requer treinamento especializado e experiência clínica, o que limita sua aplicação na prática clínica e no cotidiano das famílias, frente a identificação de possíveis comprometimentos.

Considerando esse contexto, alguns pesquisadores têm recentemente investigado o uso de métodos de aprendizado de máquina (AM) para detectar movimentos gerais atípicos de bebês. Os algoritmos empregados utilizam dados de diversos sensores como acelerômetro, sensor eletromagnético, dentre outros. Segundo a literatura, modelos treinados com dados capturados via vídeo apresentam melhor desempenho quando comparados com modelos treinados com dados obtidos por meio de outros sensores [Raghuram et al. 2021]. Isso ocorre principalmente devido ao fato de vídeos permitirem captura da profundidade dos movimentos, possibilitando que o modelo aprenda uma representação volumétrica e espaço-temporal dos movimentos, descrevendo-os de maneira mais precisa em um espaço de 3 dimensões (3D). Além disso, vídeos são uma opção não invasiva, de baixo custo, portátil e de fácil utilização nos mais diversos ambientes.

No entanto, a maioria dos trabalhos que se concentram na avaliação de GMs a

partir de vídeos utiliza modelos de aprendizado de máquina (AM) que tratam os dados em 2D, limitando o uso das informações espaço-temporais [Raghuram et al. 2022, Tsuji et al. 2020]. Os modelos de AM profundo, que são atualmente o estado na arte em reconhecimento de gestos e ações a partir de vídeo, são ferramentas que permitem tratar informações espaço-temporais. Porém, esses modelos demandam grandes bases de dados de treinamento para alcançarem altas taxas de generalização. Infelizmente, há um número muito reduzido de conjuntos de dados públicos disponíveis para reconhecimento de movimentos de bebês. Por exemplo, a base de dados apresentada em [Hesse et al. 2018] é composta por 12 instâncias de 1000 quadros gerados sinteticamente de bebês entre 4 e 6 meses. Além da quantidade limitada de instâncias, a idade dos bebês os enquadra no FMs. Por essa razão, a maioria dos trabalhos utiliza bases de dados própria, e.g. [Raghuram et al. 2022].

Neste artigo, nós propomos o uso de fusão de dados obtidos a partir de vídeos de bebês para a classificar WMs em duas classes: típico e atípico. Para explorar amplamente as informações volumétricas e espaço-temporais são utilizados dois canais de dados: 1) os próprios quadros dos vídeos RGB e 2) pontos-chaves extraídos quadro-a-quadro dos vídeos. Os dados de cada canal são utilizados para treinar uma Rede de Convolução 3D (CNN—*Convolutional Neural Network*) e respectiva rede de classificação. Portanto, são criados dois classificadores, os quais são combinados por meio de uma função de fusão. Neste trabalho, três diferentes funções de fusão são investigadas.

Os experimentos foram realizados em uma base de dados própria, composta por vídeos de bebês realizando movimentos típicos e atípicos. Embora a base de dados seja limitada em tamanho, ela contém diferentes tipos de WMs atípicos como repertório pobre, GMs caóticos, limitados, entre outros. Para melhorar os resultados, foi utilizada a estratégia de aumento de dados. Os resultados obtidos apresentam taxas de predição elevadas, considerando o tamanho limitado da base de dados. Além disso, a estratégia de fusão de dados proposta alcança taxas de classificação ligeiramente superiores às obtidas pelos modelos treinados com cada canal individualmente.

O restante deste trabalho está organizado da seguinte forma. Na Seção 2 são descritos alguns trabalhos relacionados. A metodologia proposta, bem como detalhes da base de dados criada, são apresentados na Seção 3. Os resultados dos experimentos são destacados na Seção 4. Por fim, a Seção 5 discute as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção são abordados trabalhos que lidam com o problema de avaliação dos GMs com foco na automação desse processo. A necessidade de automação surge em razão da escassez de profissionais especialistas na área, especialmente por tratar-se de uma área muito vasta. Todos os trabalhos descritos têm o objetivo principal de classificar os GMs de bebês prematuros (< 31 semanas) ou abaixo do peso, a fim de detectar precocemente quaisquer alterações neuromotoras e comprometimentos de neurodesenvolvimento por meio da identificação de atipicidade na execução dos GMs.

Uma revisão sistemática da literatura sobre tecnologias utilizadas para a automatização de GMA é apresentada em [Raghuram et al. 2021], visando analisar a acurácia do diagnóstico dessas tecnologias na predição de paralisia cerebral por meio da análise de GMs. A revisão se concentra principalmente em estudos baseados em vídeos

2D e em tecnologias de marcação 3D, nas quais os movimentos são detectados por meio de marcadores colocados no pele e capturados por sistemas de câmeras. De acordo com os autores, as abordagens que utilizam informações 3D apresentam melhor desempenho quando comparadas às abordagens que usam dados 2D.

O trabalho apresentado em [Leo et al. 2022] também oferece uma revisão geral sobre métodos de visão computacional e AM aplicados na tarefa de automação de GMA. Segundo os autores, o uso de câmeras é mais atrativo do que sensores que precisam ser fixados nos corpos dos bebês, já que esses sensores podem afetar os movimentos naturais dos bebês. Por essa razão, abordagens que utilizam dados de vídeos obtidos por câmeras ou dispositivos de profundidade são mais adequadas. Além disso, os autores citam a utilização de *pontos-chaves* como uma técnica útil para extrair atributos e analisar movimentos. Tanto vídeos RGB quanto *pontos-chaves* foram usados em nosso trabalho para aprimorar a análise de GMs.

Uma tentativa de automatizar a análise de GMs foi realizada em [Raghuram et al. 2022] com o objetivo de detectar precocemente paralisia cerebral. O estudo consistiu no desenvolvimento de um modelo estatístico baseado em vídeos 2D. Para auxiliar na extração de atributos de movimentos foi utilizada a técnica de fluxo ótico, enquanto um regressor logístico foi empregado na detecção de paralisia cerebral. Porém, os autores afirmam que o modelo não consegue detectar os GMs com elevada sensibilidade, possivelmente devido à incapacidade de capturar precisamente movimentos em 3D. O uso das informações 3D diretamente pode amenizar esse problema.

Semelhantemente à abordagem anterior, o trabalho apresentado em [Tsuji et al. 2020] propõe um sistema para avaliação dos GMs por meio de vídeos. Os autores partem do princípio que o uso de sensores vestíveis pode atrapalhar as movimentações espontâneas dos bebês. O processo de extração de atributos envolve inicialmente a remoção do fundo e a divisão dos quadros em 4 quadros. Para a classificação dos movimentos é utilizada uma rede neural chamada *log-linearized Gaussian mixture network* (LLGMN), que é composta por uma mistura de modelos gaussianos de forma logaritmo-linear, a qual estima a distribuição probabilística da base de dados.

Conforme dito anteriormente, neste trabalho nós também utilizamos vídeos e um modelo de rede neural profunda. Porém, diferentemente do que foi feito nos trabalhos descritos nesta seção, o modelo utilizado é uma CNN 3D. Além disso, nós empregamos fusão de dois canais de dados: vídeos RGB e pontos-chaves, a fim de obter resultados com maior acurácia ao considerar informações volumétricas, de movimento e espaço-temporais. A metodologia empregada é detalhada na próxima seção.

3. Metodologia

Nesta seção, nós apresentamos a metodologia utilizada para alcançar os objetivos desta pesquisa. Nós iniciamos com uma descrição detalhada da estratégia de classificação baseada em fusão empregada e da arquitetura da CNN 3D utilizada no conjunto de classificadores combinados. Em seguida, o conjunto de dados próprio utilizado é apresentado, bem como é feita a descrição das duas estratégias de representação de dados empregadas para preparar os dados para uso dos modelos de AM. Por fim, discutimos as métricas de avaliação usadas para medir o desempenho do modelo e os processos analisados visando sua otimização.

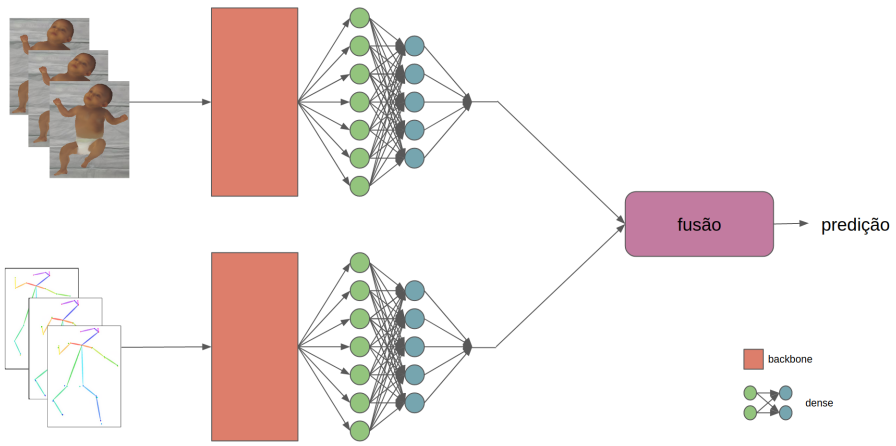

Figura 1. Estratégia de fusão de decisão dos dois modelos treinados com dados de vídeos RGB e de pontos-chaves.

3.1. Modelo de Fusão em Nível de Decisão

A Figura 1 mostra o fluxo de etapas seguido na estratégia de fusão. Primeiramente, dois canais de dados são fornecidos na entrada: 1) quadros de vídeo RGB; 2) pontos-chaves dos movimentos dos bebês por quadro. Em seguida, uma CNN 3D (*backbone*) é empregada para extrair características espaço-temporais a partir dos dados de cada canal. Cada um dos dois grupos de características extraídas é utilizado para treinar um classificador (camadas densas). A saída dos dois classificadores é combinada por meio de uma função de fusão, a qual determina a predição das instâncias.

Esse tipo de método é conhecido na literatura como abordagem de fusão em nível de decisão. Nessa abordagem, os atributos são extraídos de cada canal de dados e cada vetor de características é usado por um classificador. Assim, cada classificador atribui predições baseadas em cada canal de entrada. Em seguida, as predições individuais são agrupadas no módulo de fusão de decisões usando uma função de agregação como média, produto, etc. Por fim, a predição é feita a partir do resultado obtido pela fusão das predições individuais. Cada componente dessa estratégia é descrito a seguir.

3.1.1. Canais de Representação dos Dados

Neste trabalho nós utilizamos duas formas de representação dos dados de entrada: vídeos RGB e pontos-chaves. A combinação desses dois tipos de dados é benéfica porque os dados RGB fornecem informações importantes sobre cor e textura, bem como alguma noção de profundidade, o que pode ajudar a distinguir diferentes tipos de movimentos dos bebês. Por sua vez, os pontos-chaves fornecem informações sobre a pose e a movimentação corporal, permitindo que o classificador detecte e classifique as ações com base em padrões específicos de movimento corporal. Portanto, ao considerarmos dados de RGB e de pontos-chaves em conjunto, a expectativa é ter uma visão mais completa das informações de movimento no vídeo, permitindo uma análise mais precisa e robusta de ações em diferentes cenários. Além disso, essa estratégia permite a complementariedade

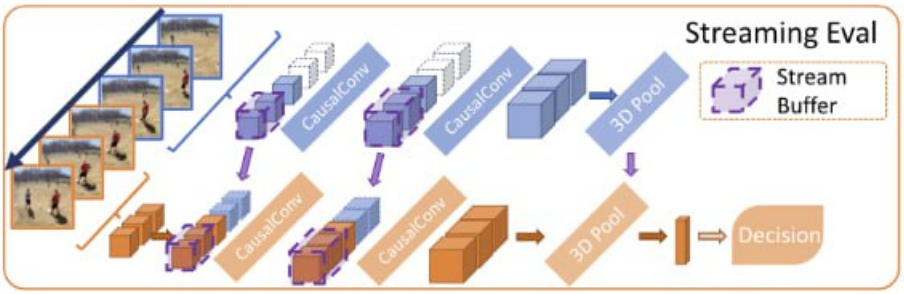

Figura 2. Utilização do *Stream Buffer* no modelo MoViNet A2 Stream. Fonte: [Kondratyuk et al. 2021]

das duas representações de dados.

Dois canais de dados foram gerados a partir dos vídeos originais. A representação em RGB consiste no próprio vídeo separado quadro a quadro. São 900 quadros por vídeo, sendo que cada quadro tem dimensões de 172 por 172 e três bandas de cor, ou seja, cada instância é representada por um conjunto de 900 quadros com as dimensões fornecidas acima. As instâncias são utilizadas como entrada para o modelo de classificação de dados RGB. Em relação às instâncias de pontos-chaves, estas foram geradas por meio da utilização da rede neural apresentada em [Chambers et al. 2020] para gerar conjuntos de pontos-chaves para os quadros dos vídeos, ou seja, para cada um dos 900 quadros de uma instância em RGB, há um quadro com 18 pontos-chaves correspondente.

3.1.2. Modelo de CNN 3D

O reconhecimento de ação em vídeos é uma tarefa desafiadora em visão computacional, exigindo uma abordagem cuidadosa e uma rede neural adequada para realizar a tarefa com precisão. Nesse sentido, a MoViNet [Kondratyuk et al. 2021] é uma família de redes neurais altamente eficientes e precisas, projetadas especificamente para o reconhecimento de ação em vídeos e padrões de movimento. Essas redes apresentam normalmente precisão elevada em comparação com outras redes neurais existentes para essa tarefa [Kondratyuk et al. 2021]. Além disso, como são modelos eficientes, podem ser executadas em dispositivos com recursos limitados, como *smartphones*. Na estratégia de fusão empregada neste artigo, nós utilizamos a versão MoViNet A2 Stream, pré-treinada.

MoViNet A2 Stream foi escolhida por ser mais eficiente em comparação com outros modelos da família MoViNet. Além de eficiente, esse modelo produz elevadas taxas de classificação, alcançando um bom equilíbrio entre eficiência e precisão por meio do uso da técnica de *Stream Buffer*, ilustrada na Figura 2. Essa técnica tem como ideia principal armazenar um pequeno buffer de quadros anteriores do fluxo de vídeo e usá-lo como entrada adicional para a rede neural que realiza o reconhecimento de ação, permitindo que o modelo leve em consideração a variação temporal nos padrões de movimento, mesmo quando há uma grande variação entre quadros individuais. É importante mencionar que os mesmos hiperparâmetros foram utilizados pelos dois modelos MoViNet A2 Stream combinados em nossa estratégia de fusão.

3.1.3. Funções de Fusão

Três diferentes funções de fusão foram utilizadas neste trabalho: produto, soma e média. Essas funções são definidas a seguir. Considerando um conjunto de n classificadores, y_i como o rótulo da classe do i-ésimo classificador e um problema de classificação com o seguinte conjunto de rótulos: $\Omega = \{w_1, w_2, ..., w_c\}$. Em nosso caso, $n = 2$, dado que nós temos um classificador treinado com os dados de vídeos RGB e outro classificador treinado com os dados de pontos-chaves. Além disso, o número de classes c é igual a 2, uma vez que o nosso problema é binário: movimento típico e atípico. Por fim, as saídas de cada classificador são fornecidas como probabilidades de classe $P(w_k|y_i(x))$, que denota a probabilidade do rótulo da classe do exemplo x ser w_k quando o classificador atribui como saída o rótulo y_i.

- **Produto**: Essa função calcula o produto das probabilidades de classe para atribuir a x a classe com maior valor de produto. A regra do produto é calculada como:

$$prod(x) = \max_{k=1}^{c} \prod_{i=1}^{n} P(w_k|y_i(x)) \tag{1}$$

- **Soma**: Nesse caso, a classe a ser atribuída a x será a classe com maior valor de soma de probabilidades. A decisão é obtida da seguinte forma:

$$soma(x) = \max_{k=1}^{c} \sum_{i=1}^{n} P(w_k|y_i(x)) \tag{2}$$

- **Média**: Calcula a média das probabilidades, conforme equação abaixo:

$$media(x) = \max_{k=1}^{c} (\sum_{i=1}^{n} P(w_k|y_i(x)))/n \tag{3}$$

3.2. Base de Dados

Os experimentos foram realizados utilizando uma base de dados própria gerada a partir da gravação de 22 vídeos de bebês prematuros ou abaixo do peso. Os vídeos foram gerados e rotulados por um especialista na área de GMA e todos representam bebês realizando movimentos do tipo WM. Como a quantidade original de vídeos é muito pequena para o treinamento de um modelo de aprendizagem profunda, algumas operações foram aplicadas para aumentarmos a quantidade de instâncias.

Como os vídeos originais têm durações variadas, precisamente entre 1:30 minuto e 3 minutos, os vídeos foram segmentados em sequências de 30 segundos, com uma frequência de 30 fps (Quadros por Segundo—*Frames per Second*), resultando em 900 quadros por instância. É importante destacar que apenas as sequências de 30 segundos em que os bebês realizam movimento atípico é que foram rotulados como instâncias da classe movimento atípico. Todas as demais sequências foram rotuladas como instâncias da classe de movimento típico. A quantidade de instâncias produzidas foi 66 no total.

Após a segmentação, os dados foram divididos em duas partições: treino e teste. Seguindo protocolo recomendado na literatura, o particionamento foi feito de forma a manter todos os dados de um indivíduo em uma única partição. Portanto, os indivíduos

Tabela 1. Distribuição das instâncias nas partições de treino e de teste

	atípico (1)	típico (0)	TOTAL
Treino	21	23	44
Teste	13	9	22

cujos dados compuseram a partição de treino foram diferentes dos indivíduos representados na partição de teste. A Tabela 1 mostra a distribuição das instâncias entre as duas partições, sendo 44 segmentos de vídeos para treino e 22 para teste.

O próximo passo empregado foi aplicar as operações de aumento de dados para vídeos quadro a quadro nas instâncias de treino. As operações utilizadas foram: translação, ruído de sal, ruído de pimenta, rotação aleatória e giro horizontal[1]. Cada uso de operações de aumento de dados gerou um novo conjunto de quadros, dos quais um novo conjunto de pontos-chave também foi obtido, sendo uma representação em pontos-chave para cada quadro.

3.3. Métricas de avaliação

A avaliação dos resultados deu-se a partir de quatro métricas: acurácia, precisão, revocação e F1-score, além da análise da matriz de confusão. A acurácia mede a proporção de exemplos classificados corretamente pelo modelo. A precisão mede a proporção de exemplos classificados como positivos que são realmente positivos. A revocação mede a proporção de exemplos positivos que foram corretamente identificados pelo modelo. O F1-score é uma medida combinada da precisão e revocação, que leva em conta tanto os verdadeiros positivos quanto os falsos positivos e falsos negativos. Ao usar essas quatro métricas juntas, é possível obter uma visão mais completa do desempenho do modelo em diferentes aspectos.

4. Experimentos e Resultados

Os nossos experimentos foram divididos em duas séries. Na primeira série, cada canal de dados foi utilizado individualmente pela rede MoViNet A2 Stream. O objetivo dessa série de experimentos é aferir o resultado obtido pelo modelo treinado com dados mono-canais, a fim de compará-los aos resultados obtidos com a fusão dos modelos treinados com os dois canais. Já a segunda série de experimentos envolve o uso do modelo de fusão e a comparação entre os resultados alcançados pelas três funções de fusão empregadas.

4.1. Resultados sem Fusão de Dados

Foram realizados dois experimentos diferentes sem uso de fusão de dados. No primeiro, os dados de pontos-chaves foram usados como entrada para treinar a MoViNet A2 Stream, enquanto no segundo, os dados de entrada foram apenas quadros dos vídeos RGB. Como pode ser observado na Tabela 2, a acurácia alcançada pelo modelo treinado com pontos-chaves foi muito baixa, 59.09%, assim como a revocação de 41.67%. Ao analisarmos a distribuição dos erros, conforme mostrada na matriz de confusão exibida na Tabela 3, vemos que o modelo errou muito mais instâncias da classe atípico (classe 1). Praticamente 60% das instâncias dessa classe foram incorretamente classificadas. Esse resultado

[1]https://github.com/okankop/vidaug

não é desejável, especialmente se considerarmos que é muito mais importante classificar corretamente as instâncias da classe atípico do que as da classe típico.

Tabela 2. Resultados obtidos no experimento com pontos-chaves (%)

Acurácia	Precisão	Revocação	F1
59.09	59.09	41.67	48.87

Tabela 3. Matriz de confusão do modelo treinado somente com dados de pontos-chaves

	Predito (1)	Predito (0)	TOTAL
Real (1)	5	8	13
Real (0)	1	8	9
TOTAL	6	16	

Os resultados alcançados pelo modelo treinado somente com dados dos vídeos RGB foram bem melhores, com acurácia de 72.22% e revocação de 61.53%, que podem ser vistos na Tabela 4. É possível observar na matriz de confusão mostrada na Tabela 5 que esse aumento foi devido à maior precisão na predição das instâncias da classe atípico. Dessa vez, o modelo classificou corretamente cerca de 62% das instâncias dessa classe. Esses resultados indicam que a representação fornecida pelos pontos-chaves não é suficiente para que sejam extraídas características espaço-temporais altamente discriminativas das duas classes do nosso problema. Já as características extraídas dos vídeos RGB parecem muito mais relevantes. No intuito de tentar explorar melhor os dois canais de dados, a próxima subseção detalha os resultados obtidos com a fusão.

Tabela 4. Resultados obtidos no experimento com vídeos RGB (%)

Acurácia	Precisão	Revocação	F1
72.22	88.89	61.53	72.77

4.2. Resultados com Fusão de Dados

Na Tabela 6 estão resumidos os resultados produzidos ao utilizarmos a combinação dos dois modelos de classificação, variando a função de fusão. O melhor valor de cada métrica está destacado em negrito. Como pode ser observado nessa tabela, de forma geral, a fusão via soma obteve os melhores resultados. A taxa de F1 (76,18%) indica melhor desempenho considerando as duas classes do nosso problema: típico e atípico. Se focarmos nas classes separadamente, porém, a fusão via média foi melhor para a classe atípico (revocação de 87,50), enquanto a fusão via soma foi a mais bem sucedida na classe típico (precisão igual a 100%).

Quando comparados aos resultados obtidos pelos modelos mono-canais, a melhor estratégia de fusão foi a soma, pois alcançou 77,27% de acurácia, enquanto a taxa obtida com dados somente dos vídeos RGB foi de 72,22%. Em termos de revocação, os dois modelos obtiveram valores iguais, precisamente 61,53%. De fato, a melhoria obtida pela

Tabela 5. Matriz de confusão do modelo treinado somente com dados de vídeos RGB

	Predito (1)	Predito (0)	TOTAL
Real (1)	8	5	13
Real (0)	1	8	9
TOTAL	9	13	

fusão via soma foi na classe típico, como mostra a matriz de confusão na Tabela 7. Todas as instâncias dessa classe foram corretamente identificadas pela combinação dos dois modelos. Por outro lado, os mesmos erros foram obtidos na classe atípico.

Portanto, os resultados das duas séries de experimentos realizados mostram que a combinação de modelos treinados com dois canais de dados superou os modelos individuais. Porém, o ganho foi pequeno. A razão para esse comportamento é provavelmente o baixo desempenho do modelo treinado com pontos-chaves. Se os resultados desse modelo fossem melhores, o desempenho da fusão dos modelos seria ainda mais superior, uma vez que a fusão dos modelos é feita em nível de decisão. A etapa de extração de pontos-chaves certamente precisa de melhorias.

Tabela 6. Resultados obtidos no experimento com fusão de dados considerando as três funções de fusão (%)

	Acurácia	Precisão	Revocação	F1
Produto	63.66	46.15	85.71	59.99
Soma	**77.27**	**100**	61.53	**76.18**
Média	63.66	53.84	**87.50**	66.66

Tabela 7. Matriz de confusão do modelo treinado com dados de pontos-chaves e de vídeos RGB usando soma como função de fusão

	Predito (1)	Predito (0)	TOTAL
Real (1)	8	5	13
Real (0)	0	9	9
TOTAL	8	13	

5. Conclusões

Neste trabalho nós utilizamos a combinação de dois modelos de classificação treinados com dados obtidos a partir de vídeos de bebês para a classificação de movimentos gerais dos bebês em duas classes: típicos e atípicos. Os modelos são do tipo CNN 3D, sendo um modelo treinado com dados dos próprios quadros dos vídeos RGB e o outro com pontos-chaves extraídos quadro-a-quadro dos vídeos. Os dois classificadores são combinados por meio de uma função de fusão. Neste trabalho, três tipo de funções foram investigadas. Além disso, os experimentos foram realizados em uma base de dados própria.

Os resultados mostram uma ligeira superioridade do método de fusão quando comparado com os modelos mono-modais. Porém, esses resultados poderiam ser melhores caso o modelo treinado com os dados de pontos-chaves apresentasse melhor desempenho. Como trabalhos futuros, nós pretendemos testar diferentes técnicas de extração

de pontos-chaves para melhorar a capacidade de representação desse canal. Também planejamos usar fusão em nível de características para avaliarmos melhor o impacto das características de cada canal na decisão final do conjunto de classificadores.

6. Agradecimentos

O presente trabalho é decorrente do projeto de Pesquisa e Desenvolvimento (P&D) 001/2020, firmado entre a Fundação da Universidade do Amazonas e FAEPI, que conta com financiamento da Samsung, usando recursos da Lei de Informática para a Amazônia Ocidental (Lei Federal nº 8.387/1991), estando sua divulgação de acordo com o previsto no artigo 39.º do Decreto nº 10.521/2020.

Referências

Aizawa, C. Y. P., Einspieler, C., Genovesi, F. F., Ibidi, S. M., and Hasue, R. H. (2021). The general movement checklist: A guide to the assessment of general movements during preterm and term age. *Jornal de Pediatria*, 97(J. Pediatr. (Rio J.), 2021 97(4)):445–452.

Chambers, C., Seethapathi, N., Saluja, R., Loeb, H., Pierce, S. R., Bogen, D. K., Prosser, L., Johnson, M. J., and Kording, K. P. (2020). Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442.

Einspieler, C., Peharz, R., and Marschik, P. (2016). Fidgety movements – tiny in appearance, but huge in impact. *Jornal de Pediatria (Versão em Português)*, 92:S64–S70.

Hesse, N., Bodensteiner, C., Arens, M., Hofmann, U. G., Weinberger, R., and Schroeder, A. S. (2018). Computer vision for medical infant motion analysis: State of the art and RGB-D data set. In *Computer Vision - ECCV 2018 Workshops*. Springer International Publishing.

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., and Gong, B. (2021). Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16020–16030.

Köpüklü, O. (2021). Video augmentation techniques for deep learning.

Leo, M., Bernava, G. M., Carcagnì, P., and Distante, C. (2022). Video-based automatic baby motion analysis for early neurological disorder diagnosis: state of the art and future directions. *Sensors*, 22(3):866.

Raghuram, K., Orlandi, S., Church, P., Chau, T., Uleryk, E., Pechlivanoglou, P., and Shah, V. (2021). Automated movement recognition to predict motor impairment in high-risk infants: a systematic review of diagnostic test accuracy and meta-analysis. *Developmental Medicine & Child Neurology*, 63(6):637–648.

Raghuram, K., Orlandi, S., Church, P., Luther, M., Kiss, A., and Shah, V. (2022). Automated movement analysis to predict cerebral palsy in very preterm infants: An ambispective cohort study. *Children*, 9(6):843.

Tsuji, T., Nakashima, S., Hayashi, H., Soh, Z., Furui, A., Shibanoki, T., Shima, K., and Shimatani, K. (2020). Markerless measurement and evaluation of general movements in infants. *Scientific reports*, 10(1):1–13.