Relatório Trabalho Prático - Aprendizagem Profunda

Grupo 5

João Barroso - PG57554
Lucas Oliveira - PG57886
Maurício Pereira - PG55984
Rafael Gomes - PG56000

Abstract. Este trabalho propõe uma abordagem baseada em técnicas de Aprendizagem Profunda para a análise automática de vídeos de treino em sutura cirúrgica, no contexto do desafio *Open Suturing Skills* (*OSS*) da *Endoscopic Vision Challenge 2025*. A solução desenvolvida integra redes convolucionais com *LSTMs* para a segmentação temporal e avaliação qualitativa de procedimentos, e aplica o modelo *Keypoint R-CNN* para a extração de características cinemáticas relevantes. Foram utilizadas diferentes estratégias de processamento de dados e anotação manual para suportar as tarefas definidas nos desafios. Apesar das limitações impostas por recursos computacionais e pela escassez de dados, os resultados mostram o potencial da abordagem proposta na avaliação automática de competências cirúrgicas.

Keywords: *Deep Learning* · Cirurgia Assistida por Computador · Vídeos Médicos · Segmentação Temporal · *OSS Challenge* · *CNN* · *LSTM* · *Keypoint R-CNN*

1 Introdução

Nos últimos anos, a integração de técnicas de visão por computador na medicina tem ganho destaque, especialmente no contexto de treino cirúrgico. Esta evolução permite não só a análise automatizada de procedimentos como também a disponibilização de *feedback* **quantitativo** e **objetivo** aos alunos. Com o crescente uso de vídeos em ambientes de simulação, torna-se cada vez mais relevante desenvolver sistemas inteligentes capazes de interpretar essas gravações e identificar padrões técnicos, erros e níveis de competência.

1.1 Contextualização e Motivação

A avaliação objetiva de competências técnicas é essencial para garantir a segurança e a eficácia dos procedimentos minimamente invasivos. Tradicionalmente, esta avaliação é realizada de forma manual e subjetiva, o que a torna morosa e suscetível a variações entre avaliadores. Com o avanço das técnicas de Aprendizagem Profunda (*Deep Learning*) e o aumento da disponibilidade de dados, torna-se possível automatizar este processo com maior robustez e escalabilidade.

1.2 Objetivos e Estrutura do Relatório

Neste contexto, o presente trabalho aborda o desafio ***Open Suturing Skills (OSS) da competição Endoscopic Vision Challenge 2025*** [1], cujo objetivo é desenvolver sistemas de inteligência artificial capazes de analisar automaticamente vídeos de treino em sutura, identificando fases do procedimento, erros técnicos e métricas de desempenho. O principal objetivo deste projeto foi conceber uma solução baseada em *Deep Learning* para a **segmentação temporal das fases da sutura**, **deteção de erros técnicos** e **extração de métricas qualitativas** a partir da análise de vídeo. Para tal, foram desenvolvidos modelos baseados em arquiteturas *CNN + LSTM* e *Keypoint R-CNN*, aplicados sobre dados de treino processados a partir de vídeos e anotações manuais.
Este relatório está estruturado da seguinte forma: na *Secção 2* é apresentada a metodologia seguida, desde o tratamento inicial dos dados até à definição das tarefas e escolha dos modelos; a *Secção 3* descreve a análise e exploração dos dados, incluindo o pré-processamento do ficheiro `OSATS.csv`, a extração de frames dos vídeos e a anotação manual de keypoints; nas *Secções 4, 5 e 6* são detalhados os modelos desenvolvidos para cada uma das tarefas do desafio *OSS* — respetivamente, *Task 1 (GRS)*, *Task 2 (OSATS)* e *Task 3 (Keypoints)*; a *Secção 7* apresenta os resultados obtidos e a análise crítica dos mesmos; por fim, a *Secção 8* discute as conclusões do trabalho e propõe direções futuras para a continuação do projeto.

2 Metodologia

A metodologia seguida neste trabalho foi desenhada com o objetivo de responder aos requisitos do desafio ***Open Suturing Skills* (*OSS*)**, integrando técnicas de Aprendizagem Profunda para análise automática de vídeos de treino cirúrgico. O processo foi dividido em três etapas principais: preparação dos dados, definição das tarefas e seleção dos modelos.
Numa fase inicial, procedeu-se à análise do *dataset* de treino disponibilizado, que incluía vídeos de procedimentos de sutura acompanhados de ficheiros de avaliação.
Em termos de definição das tarefas, a *Task 1* consistiu na classificação do desempenho global (*GRS*), enquanto a *Task 2* focou-se em múltiplas métricas específicas (*OSATS*). A *Task 3* centrou-se na extração de características cinemáticas a partir dos *keypoints* anotados, com o objetivo de relacionar variáveis como fluidez, precisão e controlo com as métricas da *Task 2*. Com base nas limitações computacionais e na análise de viabilidade, foram descartados modelos mais pesados como *I3D* e *TimeSformer*. Optou-se por arquiteturas mais leves e estáveis, nomeadamente uma combinação de *CNN + LSTM* para as *Tasks 1* e *2*, e o modelo *Keypoint R-CNN* para a *Task 3*. A escolha destas arquiteturas teve em consideração a capacidade de capturar tanto a dimensão espacial (via *CNNs*) como temporal (via *LSTMs*), bem como a possibilidade de extrair informações anatómicas e funcionais dos movimentos com o uso de *keypoints*.
A validação dos modelos foi feita com recurso a uma divisão estratificada dos

dados (*70%* treino, *30%* teste), garantindo a representatividade das classes em ambas as partes. Durante o treino, foi dada especial atenção ao balanceamento das classes através da ponderação da *loss function*. Todos os testes e desenvolvimentos foram realizados em notebooks *Jupyter*, utilizando a *framework PyTorch*.

3 Análise e Exploração dos Dados

Nesta fase inicial do trabalho, foi realizada uma análise e preparação dos dados fundamentais para o desenvolvimento dos modelos de classificação e deteção.

3.1 Processamento do ficheiro *OSATS.csv*

O ficheiro original continha múltiplas avaliações por vídeo, associadas a diferentes avaliadores. Para simplificar e uniformizar o conjunto de dados, foram realizadas as seguintes operações:

- **Conversão das 3 avaliações por vídeo numa única média**, representando uma pontuação agregada para cada dimensão avaliada.
- **Eliminação de colunas irrelevantes para o modelo**, nomeadamente: *STUDENT*, *INVESTIGATOR*, *GROUP* e *TIME*.

Este pré-processamento teve como objetivo reduzir ruído e focar o modelo apenas nas variáveis relevantes à previsão de competências.

3.2 Extração de *frames* dos vídeos

Os vídeos foram processados de forma distinta conforme a tarefa:

- *Task 1* e *Task 2*:
  - De cada vídeo, foram extraídos 16 *frames*, no formato *[16, 3, 224, 224]*, correspondendo a sequências regulares ao longo da duração do vídeo.
- *Task 3*:
  - Foram extraídos *5 frames* por vídeo, correspondendo a 1 frame por minuto, de modo a capturar momentos representativos ao longo do procedimento.

3.3 Anotação Manual (*Task 3*)

Para a *Task 3*, foram anotados *keypoints* manualmente utilizando a ferramenta *Labelme*. Cada *frame* anotado inclui *6 keypoints*, representando a posição das mãos e das ferramentas cirúrgicas. Estas anotações serviram como base para treinar um modelo de deteção de *keypoints* (*Keypoint R-CNN*).

4 *Task 1*

Nesta tarefa, o objetivo foi desenvolver um modelo capaz de classificar vídeos de procedimentos cirúrgicos segundo o sistema *Global Rating Scale* (*GRS*), que avalia competências em ***4* níveis (classes *0* a *3*)**.

4.1 Justificação da Arquitetura

Durante a fase de estudo foram avaliadas várias arquiteturas para vídeo, incluindo:

- ***I3D / SlowFast[2]:*** rejeitadas por exigirem grande capacidade computacional;
- ***ViViT / TimeSformer[3]:*** instáveis e com treino complexo;
- ***3D CNN (C3D)[4]:*** desempenho insuficiente em vídeos longos;
- ***CNN* sem *LSTM*:** não capta dependência temporal;
- ***Transformer* puro[5]:** requer muito mais dados e é propenso a *overfitting*.

A opção escolhida foi uma arquitetura ***CNN* + *LSTM***, por equilibrar boa performance com requisitos computacionais moderados:

- A ***CNN* (*ResNet34* pré-treinada)** extrai as características espaciais dos *frames*;
- O ***LSTM*** modela a evolução temporal ao longo dos *16 frames*.

4.2 Hiperparâmetros

Componente	**Valor**
Frames por vídeo	16
Resolução dos frames	224x224 (RGB)
CNN backbone	ResNet34 (pré-treinado)
LSTM hidden size	128
Número de classes	4 (GRS classes 0 a 3)
Função de perda	CrossEntropyLoss (com pesos por classe)
Otimizador	Adam
Learning Rate	1e-4
Epochs	50
Batch Size	1
Regularização (Dropout)	0.3
Estratégia de Split	StratifiedShuffleSplit (70/30)

Table 1. Hiperparâmetros utilizados na *Task 1* (*GRS*)

4.3 Resultados

- Foram utilizados *60* vídeos no total, com:
  - *42* para treino
  - *18* para teste
- O modelo obteve uma ***accuracy de 50%***.
- O modelo obteve um ***f1-score* de *53%*** na classificação *GRS*.
- Os resultados demonstram que a arquitetura é capaz de aprender padrões temporais, embora limitada pelo tamanho reduzido do *dataset*.

5 *Task 2*

A *Task 2* teve como **objetivo classificar os vídeos com base na escala *OSATS***, composta por *8* classes que avaliam diferentes aspetos da técnica cirúrgica.

5.1 Arquitetura Utilizada

Tal como na *Task 1*, foi utilizada uma arquitetura ***CNN + LSTM***, reutilizando a mesma base de extração de características e modelação temporal:

- ***CNN (ResNet34 pré-treinada)*:** extrai características espaciais de cada frame.
- ***LSTM*:** processa a sequência de frames para capturar a progressão temporal do procedimento.
- **Camada Linear:** ajustada para *8* classes nesta tarefa, em vez das *4* usadas na *Task 1*.

Esta escolha permitiu aproveitar a estrutura anterior e adaptá-la com uma pequena modificação na camada de saída.

5.2 Hiperparâmetros

A arquitetura e a maioria dos hiperparâmetros utilizados nesta tarefa mantiveram-se iguais aos da *Task 1*. As únicas alterações foram:

- O número de classes passou de *4* (*GRS*) para ***8* (*OSATS*)**, exigindo uma adaptação da camada final do modelo.
- O número de épocas de treino foi reduzido de *50* para ***10***, por questões de tempo e capacidade computacional.

5.3 Resultados

- Foram utilizados *60* vídeos no total, com:
  - *42* para treino
  - *18* para teste
- O modelo obteve uma ***accuracy de 46%***
- O ***f1-score* final foi de *49%***, refletindo a maior complexidade da tarefa (mais classes) e possíveis limitações no volume de dados.

6 Task 3

A *Task 3* teve como objetivo principal a deteção de *keypoints* das mãos e instrumentos em vídeos de treino cirúrgico, com vista à extração de métricas cinemáticas relevantes para a avaliação do desempenho técnico.

Inicialmente, foi explorada a possibilidade de utilizar a biblioteca ***MediaPipe*** para a extração automática dos *keypoints* das mãos, uma vez que esta ferramenta oferece modelos pré-treinados otimizados para deteção rápida e precisa de *landmarks* anatómicos. No entanto, os testes revelaram que o *MediaPipe* não era capaz de reconhecer corretamente as mãos nos vídeos do *dataset*, devido ao uso de luvas cirúrgicas opacas e à presença de instrumentos, o que dificultava a segmentação e rastreamento pelos modelos padrão. Esta limitação levou à decisão de realizar a **anotação manual** dos *keypoints*, utilizando a ferramenta ***LabelMe***.

Como já referido na preparação dos dados, foram extraídos 5 frames por vídeo (um por minuto), e em cada frame foram anotados seis *keypoints* principais: duas mãos e quatro instrumentos cirúrgicos (tesoura, agulha, pinça e porta-agulha). Os dados anotados foram convertidos para o formato compatível com a *API* de treino do *PyTorch*, seguindo a estrutura do *dataset COCO*.

6.1 Modelo

```
==========================================================================================
Layer (type:depth-idx)                                  Output Shape              Param #
==========================================================================================
KeypointRCNN                                            [0, 4]                    --
├─GeneralizedRCNNTransform: 1-1                         [1, 3, 800, 800]          --
├─BackboneWithFPN: 1-2                                  [1, 256, 13, 13]          --
│    └─IntermediateLayerGetter: 2-1                     [1, 2048, 25, 25]         --
│    │    └─Conv2d: 3-1                                 [1, 64, 400, 400]         (9,408)
│    │    └─FrozenBatchNorm2d: 3-2                      [1, 64, 400, 400]         --
│    │    └─ReLU: 3-3                                   [1, 64, 400, 400]         --
│    │    └─MaxPool2d: 3-4                              [1, 64, 200, 200]         --
│    │    └─Sequential: 3-5                             [1, 256, 200, 200]        (212,992)
│    │    └─Sequential: 3-6                             [1, 512, 100, 100]        1,212,416
│    │    └─Sequential: 3-7                             [1, 1024, 50, 50]         7,077,888
│    │    └─Sequential: 3-8                             [1, 2048, 25, 25]         14,942,208
│    └─FeaturePyramidNetwork: 2-2                       [1, 256, 13, 13]          --
│    │    └─ModuleList: 3-15                            --                        (recursive)
│    │    └─ModuleList: 3-16                            --                        (recursive)
│    │    └─ModuleList: 3-15                            --                        (recursive)
│    │    └─ModuleList: 3-16                            --                        (recursive)
│    │    └─ModuleList: 3-15                            --                        (recursive)
│    │    └─ModuleList: 3-16                            --                        (recursive)
│    │    └─ModuleList: 3-15                            --                        (recursive)
│    │    └─ModuleList: 3-16                            --                        (recursive)
│    │    └─LastLevelMaxPool: 3-17                      [1, 256, 200, 200]        --
...
Forward/backward pass size (MB): 1483.85
Params size (MB): 236.27
Estimated Total Size (MB): 1720.12
==========================================================================================
```

Fig. 1. Arquitetura do modelo Keypoint R-CNN

O modelo escolhido para esta tarefa foi o ***Keypoint R-CNN***, com *backbone ResNet-50* e *Feature Pyramid Network (FPN)*, disponibilizado pela biblioteca *torchvision*. Este modelo foi selecionado pela sua robustez na deteção de múltiplas instâncias com *keypoints*, beneficiando do pré-treinamento em larga escala.

As imagens foram redimensionadas para *800×800*, e aplicou-se um threshold de confiança de *0.7* para aceitar previsões. O treino decorreu com um batch *size* de *2* e uma taxa de aprendizagem de *1e-4*, durante *10* épocas.

6.2 Resultados

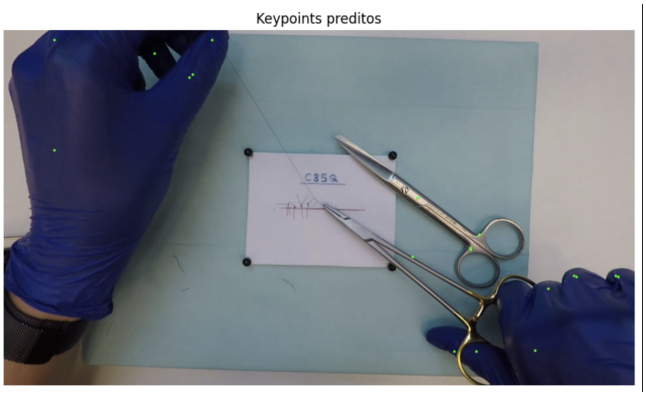

Fig. 2. Resultado obtido através da previsão do modelo

Como podemos reparar na figura 2, os resultados obtidos, embora limitados pelo número reduzido de *frames* anotados (*25* no total) no treino do modelo, mostraram-se promissores. O modelo foi capaz de detetar corretamente a maioria dos *keypoints* nas mãos e instrumentos em novos *frames*, com precisão visual aceitável e estabilidade na maioria das previsões. A performance foi avaliada de forma qualitativa, através da sobreposição visual dos *keypoints* sobre os *frames* originais. Observou-se que, em casos com menos obstrução e iluminação estável, a deteção era robusta, enquanto em cenas com movimento rápido ou sobreposição de objetos, o desempenho diminuía.

7 Resultados e Análise Crítica

7.1 Desempenho dos Modelos

Tarefa	Modelo	Accuracy	F1 Score	Observações principais
Task 1 (GRS)	CNN + LSTM	50%	53%	Boa estrutura, mas limitada por dados
Task 2 (OSATS)	CNN + LSTM	46%	49%	Tarefa mais complexa, maior número de classes
Task 3 (Keypoints)	Keypoint R-CNN	–	–	Permite extrair informações detalhadas do gesto

Table 2. Desempenho dos modelos nas três tarefas

Os modelos das *Tasks 1* e *2* conseguiram captar padrões gerais, mas foram limitados por fatores como o tamanho do *dataset* e a ausência de informações específicas sobre o gesto técnico.

7.2 Contributo Potencial da *Task 3*

Caso as *features* extraídas dos *keypoints* da *Task 3* tivessem sido integradas nos modelos de classificação (*Tasks 1 e 2*), estas poderiam ter reforçado a capacidade preditiva dos modelos, oferecendo um nível de detalhe técnico que as *CNNs* convencionais não capturam.

Feature da Task 3	Relevância para Task 1 (GRS)	Relevância para Task 2 (OSATS)
Velocidade média das mãos	Fluidez/confiança no gesto	Avaliação de motion e performance (OSATS_MOTION, OSATS_PERFORMANCE)
Tremor / Instabilidade	Indica insegurança ou má técnica	Qualidade final e fluidez (OSATS_FINALQUALITY, OSATS_FLOW)
Precisão mão-ferramenta	Reflete domínio técnico	Crucial para instrumentos e sutura (OSATS_INSTRUMENT, OSATS_SUTURE)
Tempo total de execução	Eficiência e controlo do tempo	Relacionado com fluxo e qualidade final (OSATS_FLOW, OSATS_FINALQUALITY)
Área de trabalho utilizada	Mostra foco e respeito pelo campo	Respeito e conhecimento do campo (OSATS_RESPECT, OSATS_KNOWLEDGE)

Table 3. Potencial contributo das *features* da *Task 3* nas *Tasks 1* e *2*

7.3 Análise Crítica

- A *Task 3* representa um grande valor acrescentado: mesmo não sendo usada diretamente para classificar, fornece indicadores objetivos do desempenho técnico.
- Uma possível extensão futura do trabalho seria combinar as *features* dos *keypoints* com as *embeddings* da *CNN*, criando modelos híbridos.
- Espera-se que essa fusão melhore a discriminação entre níveis de competência, especialmente em tarefas mais finas como *OSATS*.

8 Conclusão e Trabalho Futuro

8.1 Conclusão

O trabalho desenvolvido demonstrou a viabilidade de aplicar técnicas de aprendizagem profunda para a avaliação automática de competências em vídeos de treino cirúrgico. A arquitetura baseada em *CNN + LSTM* mostrou-se eficaz para modelar tanto as dimensões espaciais como temporais, resultando em desempenhos razoáveis nas *Tasks 1* e *2*:

- *Task 1* (*GRS*): *50%* de *accuracy* e *53%* de *f1-score*
- *Task 2* (*OSATS*): *46%* de *accuracy* e *49%* de *f1-score*

Apesar dos resultados estarem condicionados por restrições computacionais e limitações do *dataset*, o sistema proposto conseguiu aprender padrões relevantes relacionados com o desempenho técnico dos participantes.

A *Task 3*, baseada em deteção de *keypoints* com o modelo *Keypoint R-CNN*, permitiu extrair *features* cinemáticas como velocidade das mãos, precisão dos gestos, e área de atuação, dados valiosos que, embora não tenham sido integrados nesta fase, mostram grande potencial para reforçar os modelos de classificação.

8.2 Trabalho Futuro

Com base nos resultados e limitações identificadas, propõem-se as seguintes direções para trabalho futuro:

- Aumentar o dataset disponível e aplicar técnicas de *data augmentation* para melhorar a generalização dos modelos;
- Adicionar a métrica de *expected cost* nas *Tasks 1* e *2*, complementando as atuais métricas de *accuracy* e *f1-score*;
- Testar arquiteturas mais avançadas, como *ViViT* ou *Transformers* temporais, que podem capturar dependências de longo prazo com maior precisão;
- Combinar *embeddings* da *CNN* com features dos *keypoints*, criando um modelo híbrido multimodal mais robusto;
- Incorporar novas métricas de avaliação e técnicas de validação cruzada estratificada, aumentando a confiança nos resultados obtidos.

References

1. OSS Challenge Committee: Open Suturing Skills Challenge Wiki. https://www.synapse.org/Synapse:syn66256386/wiki/, last accessed 2025/06/02
2. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733 (2017). https://arxiv.org/abs/1705.07750
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: A Video Vision Transformer. In: International Conference on Computer Vision (ICCV), pp. 6836–6846 (2021). https://arxiv.org/abs/2103.15691
4. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks. In: IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497 (2015). https://arxiv.org/abs/1412.0767
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I.: Attention is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30 (2017). https://arxiv.org/abs/1706.03762