Video Retargeting using Vision Transformers: Utilizing deep learning for video aspect ratio change

Thesis · June 2022

DOI: 10.13140/RG.2.2.28836.35209

1 author:

Gil Laufer
KTH Royal Institute of Technology
3 PUBLICATIONS **4** CITATIONS

SEE PROFILE

KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree project in Computer Science and Engineering

Second cycle, 30 credits

Video Retargeting using Vision Transformers

Utilizing deep learning for video aspect ratio change

GIL LAUFER

TRITA – XXX-XXX 20XX:XX

Stockholm, Sweden 2022

www.kth.se

Video retargeting using Vision Transformers

Utilizing deep learning for video aspect ratio change

Gil Laufer

Master's Programme in Machine Learning

Date: 2022-06-14

Supervisor: Jonas Beskow

Examiner: Bobby L. T. Sturm

School of Electrical Engineering and Computer Science

Host company: Entecon AB

Swedish title: Video Retargeting med hjälp av Vision Transformers

Swedish subtitle: Användning av djupinlärning för ändring av videobildförhållanden

Abstract

The diversity of video material, where a video is shot and produced using a single aspect ratio, and the variety of devices that can play video with screens in different aspect ratios make video retargeting a relevant topic. The process of fitting a video filmed in one aspect ratio to a screen in another aspect ratio is called video retargeting, and the retargeted video should ideally preserve the important content and structure of the original video as well as be free of visual artifacts. Important content and important structure are vague and subjective definitions, which makes this problem more difficult to solve. The video retargeting problem has been a challenge for researchers from the computer vision, computer graphics and human-computer interaction areas, and successful retargeting can improve the viewing experience and the content's aesthetic value.

Video retargeting is done by four tools: cropping, scaling, seam carving and seam adding. Previous research showed that one of the keys to successful retargeting is to use a suitable combination of operators. This study makes use of a vision transformer, a deep learning model which is trained to discriminate between original and retargeted videos. Solving an optimization problem using beam search, the transformer assists in choosing a combination of operators that will result in the best possible retargeted video.

The retargeted videos were examined in a user A/B-test, where users had to choose their preferred variant of a video shot: the transformer's output using beam search, or a singular version where the video underwent a single retargeting operation. The model and user preferences were compared to check if the model indeed can make retargeting decisions that are appealing for humans to watch.

A significance test showed that no conclusion can be made, probably due to lack of enough test data. However, the study revealed patterns in the preferences of the users and the model that could be further fine-tuned or combined with other computer vision mechanisms in order to output better retargeted videos.

Keywords

Video retargeting, Aspect ratio, Computer vision, Deep learning, Vision transformers.

Sammanfattning

Variation av videomaterial, där olika videor är inspelade och producerade i olika bildförhållande, samt variation i apparater och skärmar som spelar upp videor i olika bildförhållanden gör ändring av videobildförhållande till en relevant fråga. Processen där en videos bildförhållande ändras heter *video retargeting*. När *video retargeting* används bör den nya videon helst bevara strukturen och viktigt innehåll från originalvideon samt vara artefaktfri. Struktur och viktigt innehåll är subjektiva definitioner vilket gör frågan svårlöst, och frågan har varit en utmaning för forskare inom datorseende, datorgrafik och människa-datorinteraktion. Lyckad ändring av en videos bildförhållande kan förbättra tittarupplevelsen och innehållets estetiska värde.

Video retargeting kan göras med hjälp av fyra funktioner: klippning, skalning, *seam carving* och *seam adding*. Tidigare studier visar att en av nycklarna till lyckad *retargeting* är att hitta en lämplig kombination av funktionerna. I denna studie används *Vision Transformer,* en djupinlärningsmodell som tränas för att skilja mellan original och omvandlade videor. Genom att lösa ett optimeringsproblem med strålsökning hjälper modellen välja den kombination av funktionerna som resulterar i den bästa möjliga omvandlade videon.

De omvandlade videorna testades genom ett användartest där användare valde vilket videoklipp de tyckte bättre om: modellens output som skapades med hjälp av strålsökning, eller en version där klippet genomgick en enklare *retargeting* med hjälp av endast en av funktionerna. Modellens och användarnas preferenser jämfördes för att se om modellen kan fatta beslut som användare upplever som bra.

Ett signifikanstest visar att ingen slutsats kan dras, förmodligen på grund av det begränsade antalet videoklipp och data som användes i studien. Däremot visar studien mönster i användarnas och modellens preferenser som kan användas för att vidareutveckla problemlösningen inom området.

Nyckelord

Video retargeting, Bildförhållande, Datorseende, Djupinlärning, Vision transformers.

Acknowledgments

I would like to thank Entecon AB and the supervisors on behalf of the company, Magnus Hoem and Jonny Lundell who hosted me for this thesis project.

Stockholm, June 2022
Gil Laufer

Table of Contents

Abstract i
Keywords i
Sammanfattning iii
Nyckelord iii
Acknowledgments v
Table of Contents vii
List of Figures ix
List of Tables x
List of Acronyms and Abbreviations xi
1 Introduction 1
1.1 Background 1
1.2 Problem 2
1.3 Goals 2
1.4 Purpose 3
1.5 Methodology 3
1.6 Delimitations 4
1.7 Outline 4
2 Background 5
2.1 Saliency Estimation 5
2.1.1 Bottom-Up Methods: Energy Maps 5
2.1.2 Top-Down Methods: Regions of Interest (ROI) 6
2.2 Retargeting Operators 7
2.2.1 Cropping 7
2.2.2 Scaling 7
2.2.3 Seam Carving 8
2.2.4 Seam Adding 10
2.3 Deep Learning 10
2.3.1 Transformers 11
2.3.2 Swin Transformers 11
2.4 Summary 12
3 Method 15
3.1 Creating Video Retargeting Operators 15
3.1.1 Cropping 15
3.1.2 Scaling 15
3.1.3 Seam Carving 15
3.1.4 Seam Adding 15
3.2 Model Development 16
3.2.1 Training Dataset Creation 16
3.2.2 Model Training 17
3.3 Video Retargeting Algorithm 17
3.3.1 Algorithm 17
3.3.2 Hyperparameters 17
3.3.3 Search Methods 18

3.4 Experimental Design 18
3.4.1 Proof of Concept 18
3.4.2 Evaluation Framework 19
4 Evaluation and Results 23
4.1 Model Evaluation 23
4.2 User Evaluation 23
4.3 Significance Test 27
4.3.1 Model-User Comparison 27
4.3.2 Paired Samples t-Test 27
5 Discussion and Analysis 28
5.1 Large Majority 28
5.2 Medium Majority 30
5.3 Small Majority 31
5.4 Model-User Decision Making 32
6 Conclusions and Future Work 35
6.1 Conclusions 35
6.2 Limitations 36
6.3 Future Work 36
6.4 Reflections 36
References 39

List of Figures

Figure 1-1:	Retargeting a 4:3 frame to 16:9 using (a) leaving black, (b) stretching and (c) cropping	**1**
Figure 2-1:	Example of an energy map	**6**
Figure 2-2:	Toy example of the dynamic programming algorithm for vertical seam carving using an energy map E (left), cumulative energy map M (center) and corresponding chosen directions (right). The chosen seam s^x is [(0,0), (0,1), (0,2), (1, 3)] which according to M minimizes the amount of energy removed from the frame.	**9**
Figure 2-3:	Comparison of seam carving choices (a – original, b – using backward energy, c – using graph cuts, d – using video energy map according to the algorithm defined in [17])	**10**
Figure 2-4:	A diagonal line (a) being distorted (c) by the removal of crossing vertical seams (b) [1]	**10**
Figure 2-5:	Video Swin-S (Small) Transformer architecture [21]	**12**
Figure 3-1:	Four original 4:3 videos from the UCF101 and their respective 16:9 retargeted videos using (a) crop, (b) scale, (3) seam carving, (d) seam adding.	**16**
Figure 3-2:	The basic retargeting search algorithm	**17**
Figure 3-3:	An original frame from “Rumble” (left) along with five retargeting operations	**19**
Figure 3-4:	Screenshot of clip B09 from the test material video. Variant A is the beam variant, a mix of cropping and scaling, while the singular variant B underwent cropping only.	**21**
Figure 4-1:	Histograms showing the distribution of the model’s scores for the beam variants (left) and singular variants (right)	**23**
Figure 4-2:	100% stacked column chart showing the votes given by the users for all 50 clips, in batch order (A01…B25) (above) and in beam vote order (below)	**24**
Figure 5-1:	Screenshots of the selected large majority clips (upper: beam variant, lower: singular variant) The variant chosen by the users is framed. If the model chose the same, the frame is green, otherwise it is red.	**29**
Figure 5-2:	Screenshots of the selected medium majority clips (upper: beam variant lower: singular variant) The variant chosen by the users is framed. If the model chose the same, the frame is green, otherwise it is red.	**30**
Figure 5-3:	Screenshots of the selected small majority clips (upper: beam variant lower: singular variant) The variant chosen by the users is framed. If the model chose the same, the frame is green, otherwise it is red.	**32**

List of Tables

Table 2-1:	Pros and cons of retargeting operators in relation to the problem definition (V – advantage, X – disadvantage, / – up to some level)	**13**
Table 3-1:	The retargeting operations and model scores for the sample clip	**19**
Table 3-2:	The shots presented in the two batches with their reference IDs and their duration (seconds:frames, 25fps)	**20**
Table 4-1:	Evaluation of the shots presented in Batch A	**25**
Table 4-2:	Evaluation of the shots presented in Batch B	**26**
Table 5-1:	Selected clips with large majority together with their beam sequence, model scores and user votes	**28**
Table 5-2:	Selected clips with medium majority together with their beam sequence, model scores and user votes	**30**
Table 5-3:	Selected clips with small majority together with their beam sequence, model scores and user votes	**31**
Table 5-4:	Min, Max and Average values of the votes given by the users and the scores given by the model for their preferred variant of each clip, according to the majority groups	**33**
Table 5-5:	Min, Max and Average values of the vote differences (users) and score differences (model) between variants of each clip	**33**

List of Acronyms and Abbreviations

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
NLP	Natural Language Processing
ROI	Regions of Interest
ViT	Vision Transformer

1 Introduction

This chapter describes the specific problem that this thesis addresses, the context of the problem, the goals of this thesis project, and outlines the structure of the thesis.

1.1 Background

In the past, video consumption alternatives were limited, mainly to the small screen, the TV device in our living rooms, or to the big screen, the cinema. These limitations led to video production in fixed aspect ratio formats such as 4:3 for TV broadcasts (where the video is 4 units wide and 3 units long) or 1.85:1 and 2.39:1, among others, for cinematic films [1]. The world of media consumption has changed drastically with the introduction of new platforms and devices, starting with the flat widescreen TVs with 16:9 aspect ratio, followed by smaller-screened laptops, smartphones and tablets that can play films and TV programs and are equipped with screens in different aspect ratios. Those mobile devices are also capable of creating content that is mostly intended for publishing in social media feeds, usually with the aspect ratio of 9:16 or 1:1, in order to fit the mobile device screens. Such content can find its way to the traditional TV broadcasts as well, for example in news reports.

The diversity of aspect ratios in which content is produced and consumed led to the need of video retargeting solutions, for adapting and resizing content from one aspect ratio to another so that its presentation would be optimal on a different platform. Traditionally, video retargeting is done in three ways (see Figure 1): stretching, leading to distortions in the proportions of the objects in the original image, cropping, leading to a risk of removing important information from the video or leaving black areas, which does not utilize the whole potential of the screen and can be problematic if the screen is small such in the case of smartphones [2].

(a) (b) (c)

Figure 1-1: Retargeting a 4:3 frame to 16:9 using (a) leaving black, (b) stretching and (c) cropping

The video retargeting problem has been a challenge for researchers from the computer vision, computer graphics and human-computer interaction areas. Solving the problem requires detection of interesting or salient areas and successful retargeting can improve the viewing experience and the content's aesthetic value [3].

In 2007, Avidan and Shamir suggested Seam Carving, an approach for solving the retargeting problem in images, where an energy map of the image is used to find seams, which are paths of pixels with minimal energy that can be removed or added for reaching the desired image size [4]. Seam Carving was later extended for use in video, where a temporal aspect is added to the equation to create

a 3D version of the image and a 2D plane is then removed from the video to achieve the size change [5]. Not being able to solve the problem completely, a big part of the research in this area is built on optimizing and improving the basic concepts of seam carving by combining them with other methods, i.e., seam carving and stretching [6] or more contemporary solutions from the world of machine learning as creating an energy map using Convolutional Neural Networks (CNNs) [7].

Recently, the use of Transformers, self-attention based deep learning models that are mainly used in Natural Language Processing (NLP) has become popular for solving computer vision tasks [19, 20]. This study aims to present a novel solution to the video retargeting problem by training a transformer to discriminate between original and retargeted videos and evaluate the process of video retargeting that will result in the most optimized retargeted video.

1.2 Problem

The core of the problem that this study aims to solve resembles the challenge of previous work in the domain of image and video retargeting [8]: Given a video V with current frame size m x n, we would like to create a new retargeted video V* with desired frame size m' x n' subject to the following constraints:

1. V* preserving the important content of V
2. V* preserving the important structure of V
3. V* is free of visual artifacts

The terms important content and important structure are vague and subjective. Due to this reason, previously found and suggested solutions that can show good results for some types of videos could not be generalized over a wide range of visual material and thus solving the problem is a nontrivial task [1].

The subjective nature of the problem makes it also difficult to evaluate retargeted videos, as there is no "ground truth" to mathematically compare a retargeted video to. It is also important to remember that while watching a retargeted video, viewers will unlikely have access to the original material but would be able to recognize anomalies based on previous knowledge and experience, for example if a person or an object would look differently than they actually do.

1.3 Goals

Given the problem definition, the goal of this degree project is to answer the following research question: *Can a deep learning-based discriminator, given an input video V, find a suitable sequence of retargeting operators that will allow the creation of a retargeted video V* which preserves the important content and structure of V and is free of visual artifacts?*

It is expected that deep learning would be able to assist in creating videos according to the constraints as described in the problem definition, up to some degree. The factors that can affect the

success rate of the deep learning model are the spatiotemporal complexity of the content present in the chosen video V as well as the desired size of V*, as a serious change in the video size will make it harder to create an optimal retargeted video.

Applying the deep learning model on a significant number of videos might help finding underlying patterns and general retargeting sequences that can provide decent results if applied on a wider range of visual material instead of a single video.

1.4 Purpose

The ability of retargeting a wide range of videos successfully is beneficial for stakeholders that make use of video material, such as over-the-top media streaming services that would like to provide programs and films produced in other aspect ratios in an appealing way for today's media consumption habits as well as archives, museums and artists that might want to use video material for research and exhibitions. Successful video retargeting can be also used by social media services for providing an enhanced user experience for users on a wider range of devices with different screen sizes.

Video retargeting has similarities to the technology used for the creation of deepfake videos, where videos are manipulated and distorted (i.e., swapping faces) in a way that makes the fake video seem authentic and hard to distinguish from real footage [14]. Although the original purpose of Seam Carving is to preserve the important content of a frame, a user can select an object in the frame to be removed by removing the seams that pass through it, creating a fake, yet realistic representation of the real footage [4]. The social issue that rises from this similarity is whether a successful method for video retargeting can be abused for creating deepfakes.

1.5 Methodology

In this study, an experimental research approach was used to answer the research question. Following the creation of the retargeting operators and training the model for discriminating between original and retargeted videos, a user test was conducted to check if the model and humans have the same preferences. The data from the model and the viewers was analyzed using statistical tools to examine the following hypothesis H_0 and the alternative hypothesis H_1:

- H_0: There is no relation between the decisions of the model and the users. Any agreements or disagreements are due to coincidence.
- H_1: There is a relation between the decisions of the model and the users. The model and the users make the same decisions (or opposite decisions).

The chosen research method is based on the following assumptions:

- The key to successful video retargeting is combining retargeting operators [12].

- The retargeting problem can be seen as a local optimization problem with an optimal solution that can be approximated using a greedy search approach.
- A deep learning model can be trained to discriminate between original and retargeted videos.

1.6 Delimitations

The project evaluates approaches and algorithm design that at a later stage could potentially be turned into a product and focuses on developing a model that will be able to solve the problem as defined in 1.2 using the existing four retargeting operators - cropping, scaling, seam carving and seam adding as they are. Hence, internal improvements to those methods will not be examined.

Due to the complexity of the problem and the required computational power for retargeting videos, the thesis project will be limited to examining one aspect ratio transformation: 2.39:1 (anamorphic widescreen format) to 16:9 (standard widescreen format). Furthermore, only shots - continuous camera recordings that end with a hard cut, fade or dissolve [1] will be examined.

1.7 Outline

The thesis is structured as follows:

- **Chapter 2** presents relevant background information about the procedures of image and video retargeting and deep learning.
- **Chapter 3** presents the realization of the retargeting operators and the development of the model used to solve the problem, as well as the creation of retargeted videos and the experiment conducted to evaluate them.
- **Chapter 4** presents the results of the experiment.
- **Chapter 5** presents a discussion and an analysis of the results.
- **Chapter 6** presents conclusions and suggestions for future work.

2 Background

Solving a visual media retargeting problem consists of two main steps [8]:

1. Identifying the salient (important) areas of the original media using adaptation algorithms.
2. Applying retargeting operators, while taking into consideration the desired target size and the information obtained from step 1.

2.1 Saliency Estimation

The first step of media retargeting is identifying the salient areas of the frame that we would like to preserve in the process. These can be areas like text, edges and other motives that can affect the human perception negatively if altered like human faces or symmetric patterns. In that way it is also possible to identify homogenous edges or motives with lower semantic content as sky or grass that can be altered, removed or multiplied without affecting the overall perception of the media [1]. This is similar to image and video compression algorithms where areas with high frequency are discarded as the human eye is less sensitive to those and so these can be removed with minimal visual effects [9]. There are two approaches for automatic detection of saliency: bottom-up and top-down [3].

2.1.1 Bottom-Up Methods: Energy Maps

Energy maps rely on traditional signal processing techniques as edge detection where filters and other mathematical calculations can be applied for finding salient areas. A simple and common energy map for an image **I** is obtained by calculating the L_1 norm gradient magnitude of every pixel and normalizing the values to [0, 1] with pixels with value 1.0 given to the most salient pixels that we would like to keep and lower value to those which can be modified [8]:

$$E(\mathbf{I}) = \left|\frac{\partial}{\partial x}\mathbf{I}\right| + \left|\frac{\partial}{\partial y}\mathbf{I}\right| \quad (1)$$

Other energy functions take into consideration other aspects of image processing and pixel-level analysis such as processing the image using a Sobel filter or detecting Harris-corners or Histogram of Gradients. No single energy function was shown to perform significantly better than others, the differences between the various functions are the parts of the frame being affected and the rate at

which artifacts are introduced [4]. Figure 2-1 presents the energy map for the image shown at Figure 1-1, with brighter areas means more energy at these points.

Figure 2-1: Example of an energy map

The energy maps for images are not sufficient for performing retargeting operations on video as changes in lighting, object positions or noise can cause introduction of serious artifacts such as jitter and blur if each frame of the video will be considered independently. Therefore, an analysis should be done on a shot (video sequence of frames shot continuously) defined as $\{I_t\}_{t=1}^{N}$ and interpreted as a 3D cube which extends the L_1 norm to a spatiotemporal L_1-norm [1, 5]:

$$E_{spatial}(i,j) = \max_{t=1}^{N} \{ \left| \frac{\partial}{\partial x} \mathrm{I}_t(\mathrm{i},\mathrm{j}) \right| + \left| \frac{\partial}{\partial y} \mathrm{I}_t(\mathrm{i},\mathrm{j}) \right| \} \tag{2}$$

$$E_{temporal}(i,j) = \max_{t=1}^{N} \{ \left| \frac{\partial}{\partial t} \mathrm{I}_t(\mathrm{i},\mathrm{j}) \right| \} \tag{3}$$

$$E_{global}(i,j) = \alpha \cdot E_{spatial} + (1-\alpha) \cdot E_{temporal} \tag{4}$$

Where $\alpha \in [0,1]$ balances the spatial and temporal energies, ideally giving more significance to the temporal energy due to the risk of introducing motion artifacts.

2.1.2 Top-Down Methods: Regions of Interest (ROI)

The human visual system does not process an image or video by analyzing it pixel by pixel but detect a Region of Interest (ROI) first, prior to further processing for object identification [10]. A different approach is therefore to use computer algorithms that detect ROIs in an image or video in a way that resembles the human vision system and define important areas that should be preserved while retargeting. Such algorithms take into consideration features as area, color, contrast, context and centrality in order to delimit ROIs and rank them in descending order of attractiveness [11].

Other popular ROI detection methods are face, text and moving object detectors. Some approaches suggest using human interaction to assist in ROI identification due to the subjectivity of the importance definition [3, 8].

2.2 Retargeting Operators

A retargeting operator is a procedure that changes a frame's width or height [12]. There are two main types of retargeting operators: discrete methods which remove or add pixels, and continuous methods which merge adjacent pixels [3].

Following the initial identification of the salient areas, the retargeting process is done by applying one or more of the following retargeting operators:

2.2.1 Cropping

Cropping is a discrete method that given a frame I with dimensions m x n creates a smaller frame I* with dimensions $m^* x\, n^*$ subject to $m^* \leq m$ and $n^* \leq n$. Given the saliency estimation of the frame, the cropping operator can solve the optimization problem of creating the new frame I* with the defined new dimensions while maximizing the presence of energy or ROI [8].

Considering the problem formulation, cropping does not induct artifacts and does preserve the structure of the frame. However, it might be impossible to preserve all important content using cropping, especially if the important content is spread in different areas of the frame, if an interest object moves outside the frame towards the cropped area or if the interesting areas are larger than the desired frame size [3]. Cropping can therefore be used as the initial retargeting operator to remove redundant information from the periphery of the frame like parts of sky or grass before moving on to other retargeting operators [1].

2.2.2 Scaling

Scaling is a continuous method where given frame I with dimensions m x n creates a new frame I* with dimensions $m^* x\, n^*$. The new size is obtained by merging pixels. The basic idea behind scaling is using a homogeneous, uniform, linear transformation to map each pixel from its current position to the new position, in either forward mapping (equation 5) or backward mapping (equation 6) [8]:

$$I(x,y) = I^*(\left\lfloor x \cdot \frac{m^*}{m} \right\rfloor, \left\lfloor y \cdot \frac{n^*}{n} \right\rfloor) \tag{5}$$

$$I^*(x,y) = I(\left\lfloor x \cdot \frac{m}{m^*} \right\rfloor, \left\lfloor y \cdot \frac{n}{n^*} \right\rfloor) \tag{6}$$

Using forward mapping cannot guarantee a one-to-one mapping which means some pixels from I could be mapped to the same new position at I^* and to empty pixels at I^*. On the other hand, backward mapping can cause losing or duplicating pixels from I [8].

Due to these issues, the main method for scaling frames is the bicubic interpolation, a process where the sixteen nearest neighbors of a point are used to estimate value of the pixel at the new point. The value of each pixel in the new frame is determined by:

$$I^*(x,y) = \sum_{i=0}^{3} \sum_{j=0}^{3} a_{ij} x^i y^j \tag{7}$$

The values of the sixteen coefficients a_{ij} are determined by solving a system of sixteen equations with sixteen unknowns using the identity at equation 6 and its derivatives I_x^* , I_y^* and I_{xy}^* [13].

Considering the problem formulation, scaling does preserve the whole content of the frame. On the other hand, it creates artifacts as blocking and aliasing and does not preserve the important structure of the frame [3]. The human vision perception is less sensitive to scaling of landscape and architecture but changes of less than 5% in the aspect ratio of faces are enough to result in an unnatural appearance. An approach that partially solves the structure problem suggests using a non-uniform transformation with higher scaling rates for the non-salient parts of the frame [1]. However, the trade-off of this approach is that it amplifies artifacts and distortions [8].

2.2.3 Seam Carving

Seam carving is a discrete method where a connected, one pixel wide, path of pixels (seam) that goes all the way from top to bottom or left to right is removed, thus decreasing the width or the length of the frame by one pixel and preserving the frame's rectangular shape. The advantage of using seam carving in comparison to cropping is that seam carving is not limited to the periphery of the frame or to straight lines. The mathematical definitions of a vertical seam $\mathbf{s^x}$ and horizontal seam $\mathbf{s^y}$ are the continuous and discrete functions x, y as follows [1, 3, 8]:

$$\boldsymbol{s^x} = \{s_i^x\}_{i=1}^n = \{(x(i), i)\}_{i=1}^n, s.t. \forall i, |x(i) - x(i-1) \leq 1 \tag{8}$$

$$\boldsymbol{s^y} = \{s_j^y\}_{j=1}^m = \{(j, y(j))\}_{j=1}^m, s.t. \forall j, |y(j) - y(j-1) \leq 1 \tag{9}$$

Using the saliency estimation of the frame, for example using the energy map obtained by equation 1, the optimal seam $\mathbf{s}^*$ with the minimal sum of energy E(s) of its pixels is removed:

$$\boldsymbol{s}^* = \arg\min_s E(s) = \arg\min_s \sum_{i=1}^{n} E(I(s_i)) \tag{10}$$

Finding the optimal seam is done by calculating backward energy using dynamic programming in linear complexity, by creating a cumulative energy map M. Beginning at the last row of the energy map E, we iterate all way up finding the best seam connectivity for each pixel (up-left, up-straight or up-right) where the sum up to that point is minimized. The sum at the matching position of M together with the chosen direction are noted. The process terminates when the first row is reached. The smallest value at the first row of M points at the end point of the seam. After the removal of the seam, the energy map is updated accordingly and the dynamic programming algorithm runs again to find a new optimized seam to be removed, until the desired new frame size is reached [8]. A toy example can be found in Figure 2-2.

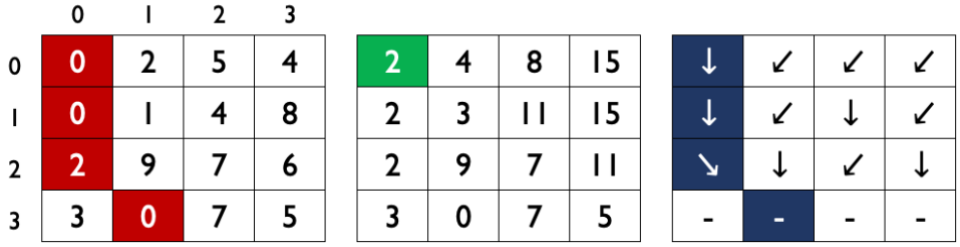


Figure 2-2: **Toy example of the dynamic programming algorithm for vertical seam carving using an energy map E (left), cumulative energy map M (center) and corresponding chosen directions (right). The chosen seam s^x is [(0,0), (0,1), (0,2), (1, 3)] which according to M minimizes the amount of energy removed from the frame.**

The backwards energy algorithm however ignores energy that is inserted to the frame due to high-energy pixels becoming neighbors following the removal of a seam, which will produce visible artifacts. A forward energy algorithm was therefore proposed instead where the chosen seam to be removed does not necessarily have the minimal energy but inserts the minimal amount of energy upon removal [8].

Extending seam carving to video requires handling a video as a 3D cube due to the added temporal dimension. A spatiotemporal energy map, for example as described in equations 2-4, can be used to find a video seam where a connected 2D manifold "surface" that cuts through the 3D model is removed. However, adding a third dimension to the dynamic programming algorithms is a complicated task and therefore graph cuts are suggested for applying seam carving on videos [5].

Each pixel of the frame or video is represented as a node with directed weighted arcs, based on the saliency estimation, connecting each pixel to its neighbors. All leftmost pixels are connected to a source node s and all rightmost pixels to a sink node t. Using a maxflow/min-cut algorithm, the target is to find a cut that will partition the graph into two sets S and T with the weight of the removed edges connecting S to T is minimum. Merging and removing edges and setting the weights of some edges to infinity ensure that the cut will comply with the definition of a seam, and that the obtained cut will allow removal of the seam with the lowest energy amount [5].

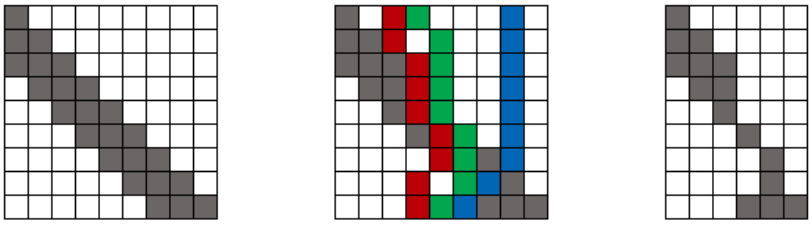
Although feasible compared to the dynamic programming algorithm, graph cuts have a high computational complexity which makes them unusable for videos due to the required memory and processing times [1]. Newer methods focus on creating a single energy map for a video, based on the theory that videos may have redundant content that does not, or slightly change from frame to frame, and so applying the same calculations multiple times is unnecessary. Calculating energy maps for each frame and averaging them to a uniform energy map allows finding and removing the same seams from all frames at the same time, which eliminates the risk of displaced content (jittering) between frames [17].

The choice of saliency estimation and carving algorithm can affect the visual result, as different choices will produce different outputs from the same input (Figure 2-3).

(a) (b) (c) (d)

Figure 2-3: Comparison of seam carving choices (a – original, b – using backward energy, c – using graph cuts, d – using video energy map according to the algorithm defined in [17])

Seam carving preserves the important information and structure of the frame but might not create artifact-free media. A common artifact is when the original frame includes patterns or straight, non-horizontal or non-vertical lines so that removing seams crossing these lines result in distortions as seen in Figure 2-4. Furthermore, in correspondence to the desired new frame size, seam carving might come to a point where the candidate seam to be removed has a significant amount of energy due to prior removal of lower-energy seams, resulting in unwanted or unnatural cuts within the frame.



(a) (b) (c)

Figure 2-4: A diagonal line (a) being distorted (c) by the removal of crossing vertical seams (b) [1]

2.2.4 Seam Adding

Seam Adding is a continuous method where a connected, one pixel wide, path of pixels (seam) that goes all the way from top to bottom or left to right is added, thus increasing the width or the length of the frame by one pixel and preserving the frame's rectangular shape. To find an amount of k seams that will be inserted with their location in the frame and pixel values, an inversion of the frame is approximated by finding the k seams that would have been removed using seam carving. By tracking the paths of the seams that were removed, it is possible to iterate on the original media and go through the same seams for adding new values along each seam. The pixel value at each new position is calculated as the average of its left and right neighbors (vertical seam) or top and bottom neighbors (horizontal seam) [4].

2.3 Deep Learning

Deep learning is a subfield that belongs to the domain of artificial intelligence and machine learning. Deep learning models are large-scale Artificial Neural Networks (ANNs) that given a complex dataset as input (the training dataset), can identify and extract patterns that are used to make data-driven

decisions on new data [16]. Deep learning applications are popular in the field of computer vision, where a neural network is trained to interpret images and videos for further prediction or decision making. Thus, deep learning can be trained to understand the content and structure of natural videos as well as the appearance of artifacts, knowledge that can be applied for finding an appropriate combination of operators for retargeting a specific video according to the problem definition.

2.3.1 Transformers

Transformers are deep learning models that use the mechanism of positional encoding, attention and self-attention to process input and find a suitable output. The transformer was first developed as an alternative for machine translation in NLP, translating English to German or English to French [18, 20]. While training for processing text with sentences in English and their equivalent translations in French, the words, represented as vectors, along with their positions in the sentence, are sent as separate data which makes it easier to train larger amounts of data in parallel. The attention mechanism helps the transformer understand which words in the input data are more correlated to words in the output data and therefore should be noticed, while the self-attention mechanism is doing the same on the input alone. The advantage of transformers, apart from being efficient to train, is their ability to long-range dependencies in the data.

Applying the same mechanism in computer vision became possible with the introduction of Vision Transformers (ViT) [19]. ViT make use of the original transformer as closely as possible: as the original transformer handles sentences which are vector representations of words in a specific order, an image is represented as a sequence by splitting the image to linearly projected 16x16 patches with noted positions. Experiments showed that ViT work well when pre-trained on large datasets and then fine-tuned for smaller recognition tasks. ViT showed promising results on image classification which requires a single output for a whole image but other computer vision tasks as detection and segmentation, where decisions need to be made at the pixel level, remained a challenge.

2.3.2 Swin Transformers

The challenges ViT did not manage to solve were set to be solved by Swin Transformers, which more general and hierarchical transformers with **S**hifted **win**dows based self-attention [20]. The idea behind this concept is to limit the attention span to a smaller number of patches. The limited output is then merged for creating larger patches, thus making it possible for the transformer to analyze information from a larger area. The attention window is then shifted in relation to the previous layer and the process repeats itself according to the transformer's configuration or until no merges are possible.

Video Swin Transformer is an extension of the Swin Transformer to the spatiotemporal domain, where a video is partitioned into 3D patches and uses cube structures as windows [21]. The Video Swin Transformer has four main steps, with steps 3 and 4 repeating themselves in relation to the model architecture (see Figure 2-5):

1. 3D Patch Partition – the initial partition of the input to 3D patches.
2. Linear Embedding – converts each patch to a C-dimensional vector.
3. Swin Transformer Block – where the self-attention is computed.
4. Patch Merging – acts as downsampling.

Video Swin Transformer has four suggested architecture variants: Swin-T (Tiny), Swin-S (Small), Swin-B (Base) and Swin-L (Large), which differ at their C-value and the number of blocks at each layer. Swin-S, as presented in Figure 2-5, have 96 channels of hidden layers at the first stage and layer structure {2, 2, 18, 2}, and is about half of the size and complexity of the base model.

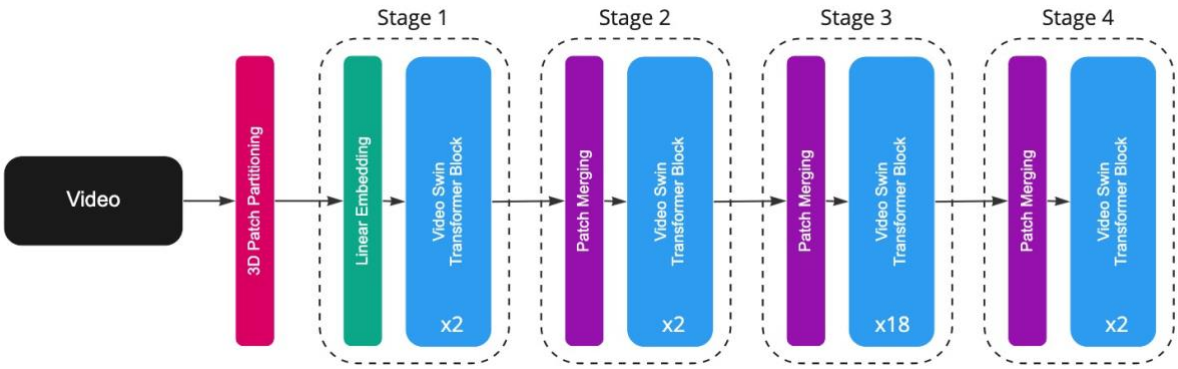


Figure 2-5: Video Swin-S (Small) Transformer architecture [21].

The input of Swin Transformers does not have to be fixed in a specific size or shape as the input is segmented and tokenized, similarly to the original NLP transformers, which can analyze sentences in every length. Videos with larger frame size or more frames will simply result in longer "sentences". A video can be inputted as a folder of sequential frames, each saved as an image file, or as a video file, and the model will do the separation to frames itself. Inputting frames instead of video can speed up the training process as the transformer won't have to extract the frames again and again.

As other neural networks, Swin Transformers learn using stochastic gradient descent, in this case with the help of Adam optimizer. Testing on Kinetics-400, a collection of video clips that cover 400 human action classes, Swin-S and Swin-L achieved 80.6% and 84.9% top-1 accuracy respectively on video action recognition [21].

2.4 Summary

This chapter covered the background for understanding media retargeting and the deep learning techniques that are required to create the video retargeting model that is evaluated in this thesis project. This chapter answered the following main questions:

- How can we find the important areas of a frame that we would like to preserve?
- What retargeting operators are available in our "toolbox"? How do they work and what are the pros and cons of each? (See a summary in Table 2-1)

- What is a transformer, and why is it suitable for solving our problem?

Table 2-1: **Pros and cons of retargeting operators in relation to the problem definition (V – advantage, X – disadvantage, / – up to some level).**

	CROPPING	SCALING	SEAM CARVING	SEAM ADDING
PRESERVES CONTENT	X	V	X	V
PRESERVES STRUCTURE	V	X	/	/
ARTIFACT-FREE	V	X	/	/

3 Method

This chapter describes the development of the components needed for the retargeting process and the algorithm that connects them all together, followed by a description of the experiment set-up and the methods used to evaluate the algorithm.

3.1 Creating Video Retargeting Operators

All retargeting operators were created for performing vertical operations only. For operating horizontal operations, the video is first transposed, and transposed back after completing the operation. The retargeting operators receive, apart from the input video, also the desired output size as parameter.

3.1.1 Cropping

Cropping was implemented by taking the central slice of the frame defined by the desired output size. For example, cropping 4% will result in removing 2% from the left and 2% from the right.

3.1.2 Scaling

Scaling was implemented using OpenCV's cv2.resize() with intercubic interpolation.

3.1.3 Seam Carving

Seam carving was implemented as an adaptation of the fast video retargeting, seam-carving based algorithm by Zhu [17]. In this algorithm, energy maps are calculated using Sobel filters for each frame in the video. The frames are batched in buffers in relation to the average standard deviation of energy, and a uniform energy map defined as the average of the energy maps of all frames in the buffer is created. The calculated seams, according to the averaged single energy map, are then removed from all seams in the buffer.

The division of the video to buffers can result in several buffers even for short clips, which can cause jitter if different seams are removed in different buffers, for example a man writing on board where the text on the board moves around. Therefore, the average standard deviation of energy is not used, and the algorithm relates to the whole video as a single buffer.

3.1.4 Seam Adding

Seam carving is implemented as an extension to seam carving. First, the desired k seams are removed using the algorithm described in 3.1.3. Their paths are tracked and then added to the original video.

3.2 Model Development

Based on the characteristics of transformers and the generic purpose of Video Swin Transformer in particular, it was assumed that such a transformer can be trained to discriminate between original and retargeted videos and should perform well using pre-trained weights that are fine-tuned for that specific recognition task.

3.2.1 Training Dataset Creation

The training data for the fine-tuning of the transformer is based on UCF101, an action recognition dataset of 13220 realistic action videos collected from YouTube, featuring 101 human actions in five categories [22]:

1. Human-Object Interaction (for example Apply Lipstick, Pizza Tossing)
2. Body-Motion Only (for example Blowing Candles, Walking with a Dog)
3. Human-Human Interaction (for example Haircut, Military Parade)
4. Playing Musical Instruments (for example Flute, Violin)
5. Sports (for example Biking, Ice Dancing)

UCF101 was chosen as a starting point thanks to its documented support with the Video Swin Transformer and its video characteristics: it includes videos in various aspect ratios, about 65% 4:3 and 35% 16:9, with each video being a few seconds long shot presenting one of the 101 activities. In this case, the original classes of the dataset were insignificant. Instead, a sample was taken and divided into two classes:

1. Original videos – 3800 clips taken from the 16:9 videos of UCF101
2. Retargeted videos – 3800 retargeted clips using a single retargeting operator: scale, crop, seam carving or seam adding, 25% of the 3800 for each operator. Examples of the retargeted videos are shown in Figure 3-1.

The 7600 clips were divided into training (80%), testing (10%) and validation (10%) datasets.

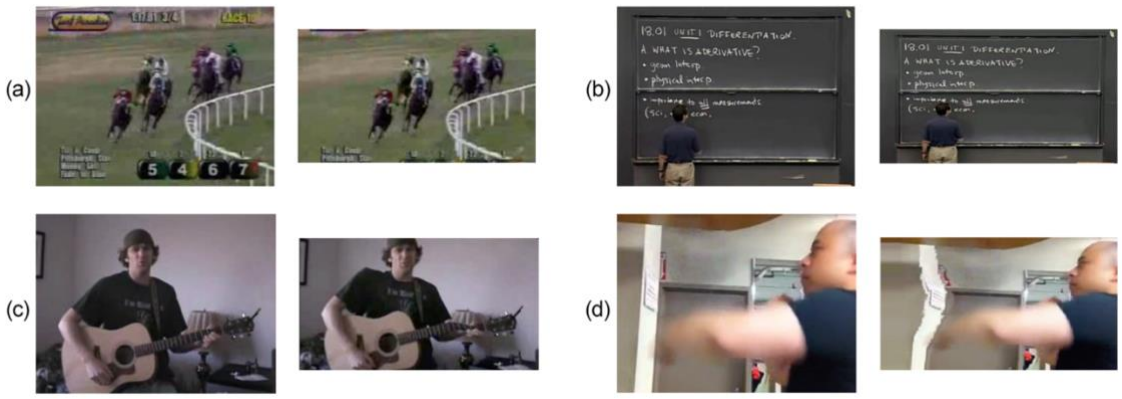


Figure 3-1: **Four original 4:3 videos from the UCF101 and their respective 16:9 retargeted videos using (a) crop, (b) scale, (3) seam carving, (d) seam adding.**

3.2.2 Model Training

The Video Swin Transformer model chosen for the task is Swin-S. The model was initiated using pre-trained weights of ImageNet-1K [27]. Fine-tuning the weights with the newly created dataset reached its maximum point after 35 epochs, discriminating successfully between original and retargeted videos in an accuracy of 95.7%. One epoch of training with the Swin-S model and the modified dataset took about 30 minutes.

3.3 Video Retargeting Algorithm

The retargeting operators and the model developed at the previous stages are used as black boxes in the overall retargeting process. Aspects need to be considered while using the algorithm are the hyperparameters and search methods as described later in this section.

3.3.1 Algorithm

The algorithm presented in Figure 3-2 is a while-loop that retargets the video to a chosen size with all retargeting operators separately and evaluates the outcomes using the transformer. The best alternative(s) according to the transformer is/are chosen for further retargeting, until the desired aspect ratio is reached.

```
while ASPECTRATIO(video) != desiredAspectRatio:
        a = CROP(video, changeRate)
        b = SCALE(video, changeRate)
        c = SEAMCARVE(video, changeRate)
        d = SEAMADD(video, changeRate)
        bestOutput = TRANSFORMER(a, b, c, d)
        video = bestOutput
```

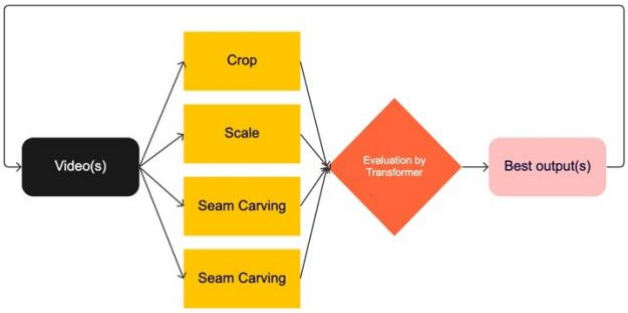


Figure 3-2: The basic retargeting search algorithm

For simplicity, a "greedy" option where only the best alternative at each step is taken is shown, but a beam search can be used where the top n best options are taken at each step, and the rest discarded. The **TRANSFORMER** returns a list of probabilities for `video` to be classified to one of the classes (original, retargeted). `bestOption` is considered the retargeted video with the highest probability to belong to the original class, even if it is less than 0.5.

3.3.2 Hyperparameters

Using the algorithm requires a series of pre-chosen hyperparameters and limitations based on the desired change (input and output aspect ratio) in order to make sure that the retargeting process

converges. Without hyperparameters, the algorithm might go towards an undesired aspect ratio or get in a loop for example by choosing an operator that enlarges the video followed by an operator that shrinks the video and so on. The hyperparameters that need to be considered while working with the algorithm are:

- Change rate: how much is added or removed at each step? Different change rates can be chosen at different stages and for different operators.
- Directionality: Making horizontal and/or vertical operations.
- Convergence: If a smaller aspect ratio is desired, enlarging operations should be limited or eliminated, and vice versa.

3.3.3 Search Methods

The basic search method where only the best (top 1) state is kept at each step, could be expanded to beam search, continuing with the top n states at each step according to the model's evaluations, until reaching the final step and choosing the best final state. Considering more partial states can potentially reach a final state with a higher model score which did not have the best score at earlier steps.

3.4 Experimental Design

The experiment focuses on changing anamorphic widescreen format (2.39:1) used in modern cinema to standard widescreen format (16:9) used in modern televisions and computer monitors [15]. The change from 2.39:1 to 16:9 (1.77:1) was done by applying a series of seven vertical operations, each reducing the video with by 4% ($2.39 \times 0.96^7 \approx 1.77$). Therefore, the scale operator was used to reduce the image size and seam adding was not used at all. The beam search had a beam size of 5, which means that a single video had to go through 87 retargeting operations before finding a complete sequence of seven operations which yielded the best result among the other candidates.

In the rest of the thesis, the retargeting operators will be referred as numbers, both standalone and in sequences, with 0 for crop, 1 for scale and 2 for seam carving. For example, A retargeting sequence of 0011122 means that the respective video has gone through two crops, followed by three scales, followed by two seam carvings.

3.4.1 Proof of Concept

First, we ran the algorithm on a sample clip taken from the trailer of the film "Rumble" where two characters stand in front of each other in a crowded stadium. The characters appear at the sides of the frame with an object (a bell) placed between them. While retargeting the video, the characters, the bell and the background should be considered. The original video was retargeted in five different ways and scored by the model, as described in Table 3-1 and seen in Figure 3-3:

Table 3-1: **The retargeting operations and model scores for the sample clip**

		SEQUENCE	SCORE			SEQUENCE	SCORE
a	Beam Search	2010000	0.435	**c**	Crop	0000000	0.298
b	Greedy Search	0102000	0.402	**d**	Scale	1111111	0.332
				e	Seam Carving	2222222	0.372

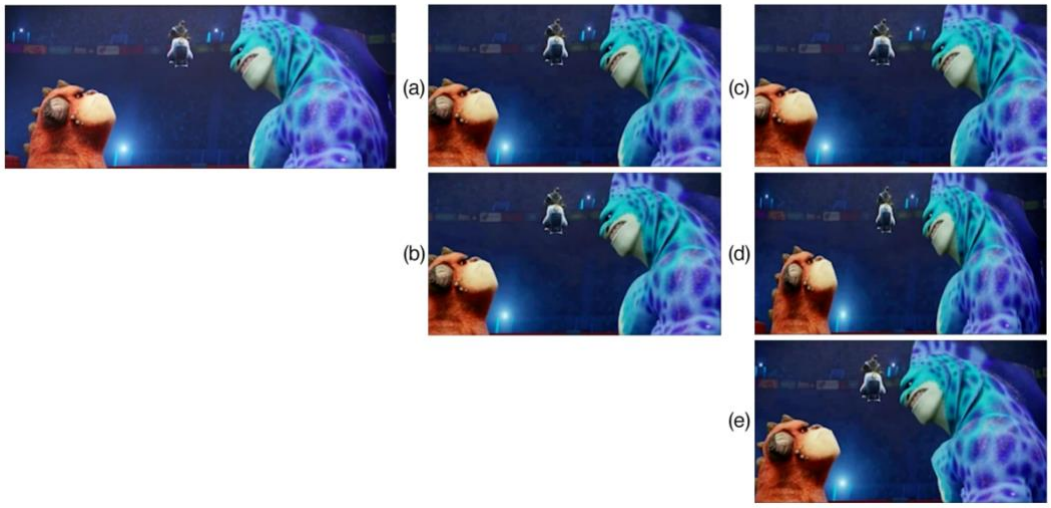


Figure 3-3: **An original frame from "Rumble" (left) along with five retargeting operations**

While crop (c) did not preserve all content of the original frame, scale (d) did not preserve the structure and seam carving (e) introduced artifacts as seen on the right character, the combination of operators managed to reach a result less extreme than the singular operations, with the model scoring the variants accordingly. This example shows that beam search has the potential of finding and identifying the best possible retargeting combination, which in turn is better than performing singular operations. Therefore, a wider experiment could show if the model indeed assists in performing retargeting operations while choosing the most appealing visual result.

3.4.2 Evaluation Framework

The model was evaluated by retargeting shots of film trailers from 2.39:1 aspect ratio to 16:9. A total of 50 shots were extracted from 35 trailers using PySceneDetect, a command-line interface, Python based package that detects scenes in a video and splits the trailer to smaller shot videos. As film trailers usually have many quick cuts, the chosen shots had to be at least three seconds long or two seconds long in case of a static camera to make sure that the participants of the test will manage to get an understanding of the videos. To ensure quicker operations while preserving image quality, the trailers were re-encoded from HD quality to 576p.

The 50 shots were processed in two retargeting tracks, creating two variants for each shot:

1. Beam search: choosing the candidate video with the highest model score presented by the model (beam score) after the seven operations.

2. Singular search: choosing the candidate video with the highest model score presented by the model (singular score) after seven operations of the same operator.

The shots were presented in a user A/B test, where the user had to decide which variant of each shot looked better, with the purpose of finding out whether users agree or disagree with the model's decisions and preferences. Two batches of shots A and B were created with 25 videos in each (see Table 3-2), divided in a way that the groups will have as similar mean value and standard deviation of beam scores as possible. The presentation order in each group was assigned randomly.

Table 3-2: **The shots presented in the two batches with their reference IDs and their duration (seconds:frames, 25fps)**

ID	FILM	DURATION	ID	FILM	DURATION
A01	The Neon Demon	06:17	**B01**	Hundraåringen	06:20
A02	The Neon Demon	02:13	**B02**	The Sitter	04:13
A03	The Tomorrow Man	02:22	**B03**	The Tomorrow Man	02:22
A04	Corn Island	08:01	**B04**	Annie	03:22
A05	3 Days With Dad	04:10	**B05**	The Tomorrow Man	03:09
A06	Finding Your Feet	04:24	**B06**	Old	03:09
A07	Mandy	04:15	**B07**	Kung Fu Panda	03:21
A08	Darjeeling Limited	08:06	**B08**	Hollow Point	03:10
A09	Devil	06:07	**B09**	Cars	03:01
A10	Thoroughbreds	03:06	**B10**	Tower Heist	04:22
A11	Allies	04:19	**B11**	Corn Island	04:02
A12	The Sitter	04:02	**B12**	The Secret Sculpture	02:16
A13	The Shack	02:00	**B13**	XX	03:14
A14	JP Ja Murtovaras	02:14	**B14**	The Neon Demon	04:21
A15	Aluna	03:19	**B15**	Air Force One	03:21
A16	The Sitter	04:12	**B16**	The Sitter	04:14
A17	Dawn	04:12	**B17**	Monsters vs Aliens	05:20
A18	Thoroughbreds	04:01	**B18**	Finding Your Feet	09:13
A19	Tron	02:12	**B19**	The Reason I Jump	03:06
A20	Black Mass	04:12	**B20**	Lamb	07:03
A21	The Reason I Jump	03:13	**B21**	Julenatt i Blåfjell	02:12
A22	Mayday	04:24	**B22**	Rivers Run Red	04:09
A23	Darjeeling Limited	04:15	**B23**	The Tomorrow Man	02:20
A24	Breathless	04:21	**B24**	Monsters vs Aliens	04:22
A25	Daughter of The Wolf	01:17	**B25**	Cars	02:02

Each user was assigned a batch and received a video file including all 25 clips. Each shot was presented three times in row. The changes in the video were only vertical, therefore the clips were placed one above the other, to make it easier to examine the differences between the variants, as seen in Figure 3-4. The location of each variant (top/bottom) was assigned randomly for each clip.

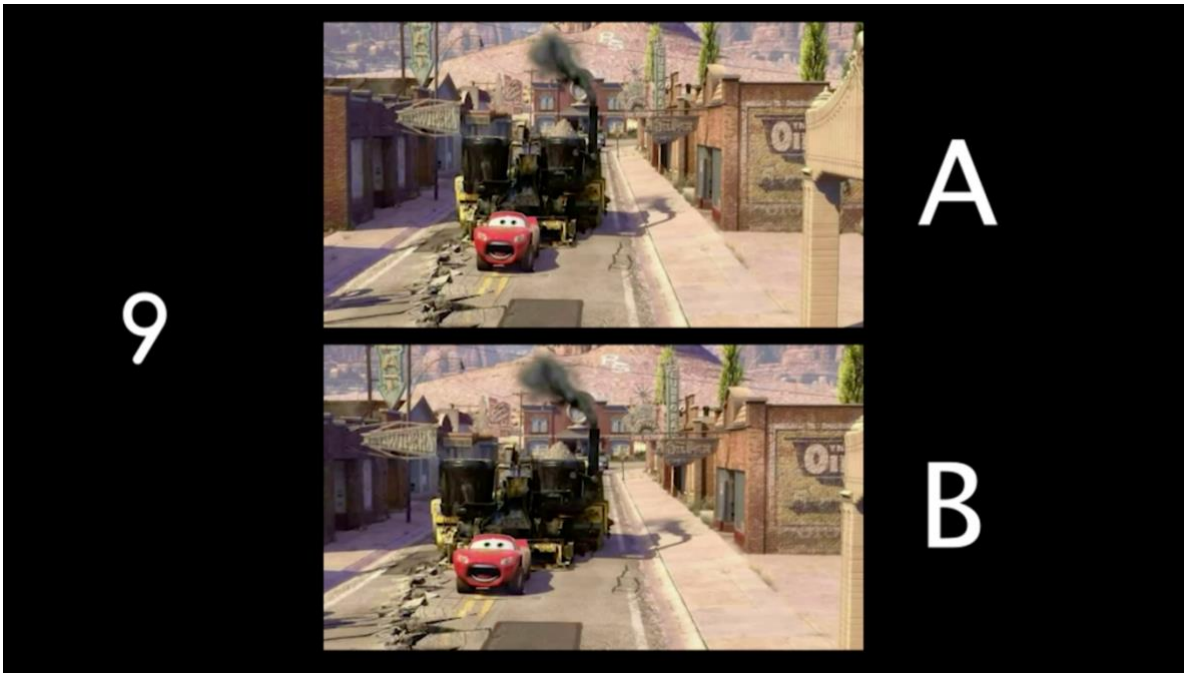


Figure 3-4: **Screenshot of clip B09 from the test material video. Variant A is the beam variant, a mix of cropping and scaling, while the singular variant B underwent cropping only.**

4 Evaluation and Results

In this chapter, the results of the evaluation by the model and the users are presented and examined using a statistical paired samples t-test. Tables 4-1 and 4-2 presents an overview of the 50 clips along with their beam sequence, singular operator, model's scores, user votes and the difference between the preferred variants, which is necessary for the statistical significance test. No difference is marked as 0, model choosing beam and users choosing singular is marked as 1, and vice versa as -1.

4.1 Model Evaluation

The model examined each clip variant independently, with no relation to the other clip variant, and gave it a grade in which shows the probability that the variant is untouched (not retargeted). Figure 4-1 presents the histograms of the model's scores for the beam variants and singular variants, which shows that the model had an overall preference of beam variants. On average, the beam variants received a model score of 0.529, a score which classifies a video as not retargeted, while the singular variants received a model score of 0.477, which classifies a video as retargeted.

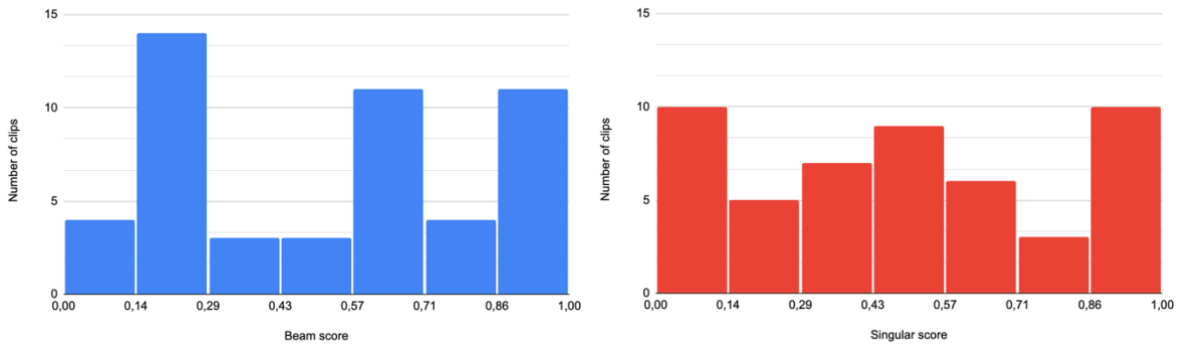


Figure 4-1: Histograms showing the distribution of the model's scores for the beam variants (left) and singular variants (right)

According to the results, the shots can be generally divided into three groups:

1. Shots with beam score larger than 0.5 and beam score larger than singular score (25 shots), marked in Tables 4-1 and 4-2 in green.
2. Shots with singular score greater than beam score, due to the beam not being wide enough (15 shots), marked in Tables 4-1 and 4-2 in red.
3. Shots with beam score smaller than 0.5 and beam score larger than singular score (10 shots), marked in Tables 4-1 and 4-2 in blue.

4.2 User Evaluation

A total of 50 users took part in the user test, males and females aged 20-50. Each user watched one of the batches so in total each batch and clip were watched by 25 users. Each user had to decide which of the variants is more appealing. An odd number of users was chosen to avoid ties and make sure that one of the variants will always have a majority. Figure 4-2 presents the distribution of the user

votes per clip, which shows a user preference towards the beam variants. On average, the beam variants received 57% of the votes while the singular variants received 43%.

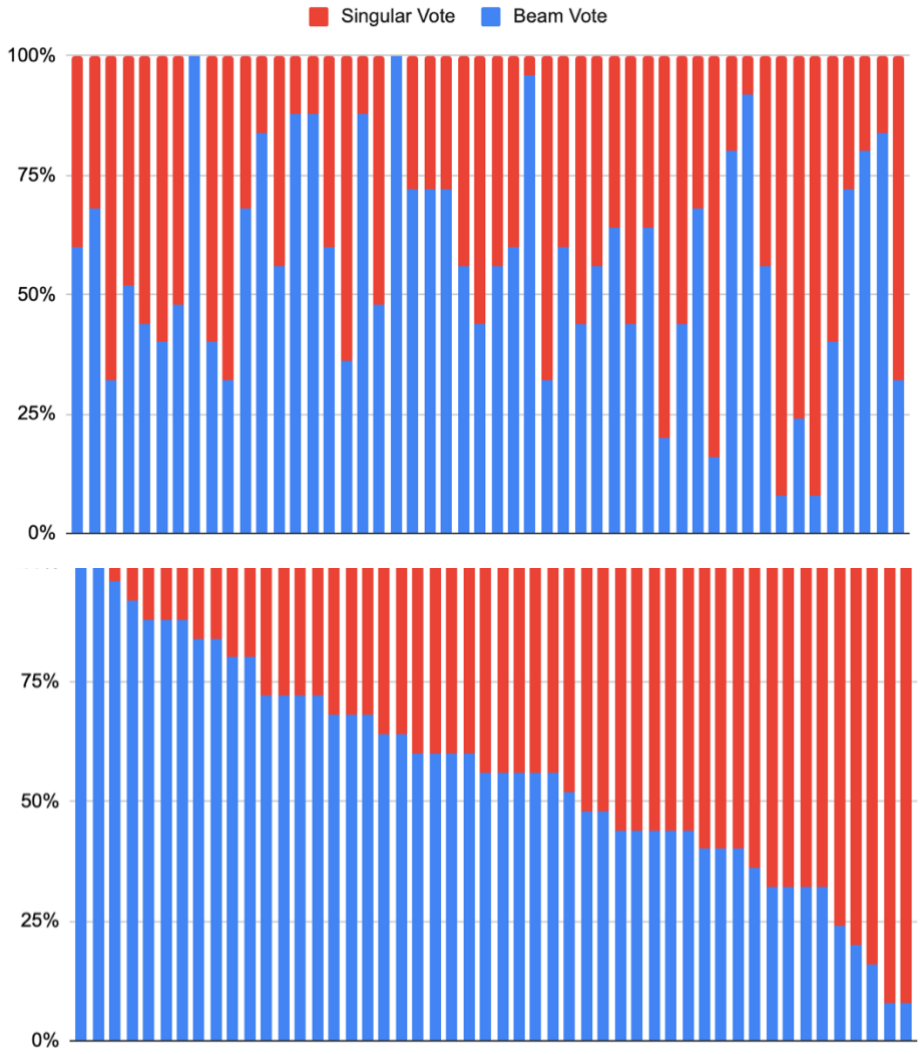


Figure 4-2: **100% stacked column chart showing the votes given by the users for all 50 clips, in batch order (A01…B25) (above) and in beam vote order (below)**

Considering that the lowest score a variant could receive for being declared as the preferred variant is 52, the shots are divided into three groups:

1. Shots with large majority vote, with the preferred shot receiving 72-100% of the votes. A total of 20 clips belong to this group.
2. Shots with medium majority vote, with the preferred shot receiving 60-68% of the votes. A total of 17 clips belong to this group.
3. Shots with small majority vote, with the preferred shot receiving 52-56% of the votes. A total of 13 clips belong to this group.

Table 4-1: Evaluation of the shots presented in Batch A

CLIP	SEQUENCE	BEAM SCORE	SING. SCORE	BEAM VOTE	SING. VOTE	DIFF
A01	0002200 / 0	0.999988	0.999986	60	40	0
A02	0011202 / 1	0.594	0.512	68	32	0
A03	0222101 / 1	0.799	0.638	32	68	1
A04	2121121 / 1	0.670	0.656	52	48	0
A05	1101111 / 1	0.009	0.020	44	56	0
A06	1110000 / 0	0.981	0.975	40	60	1
A07	1110121 / 1	0.514	0.471	48	52	1
A08	1111222 / 2	0.513	0.468	100	0	0
A09	1210111 / 1	0.232	0.179	40	60	1
A10	1020010 / 0	0.599	0.504	32	68	1
A11	0212111 / 1	0.269	0.294	68	32	-1
A12	1011102 / 1	0.754	0.767	84	16	-1
A13	1111101 / 1	0.198	0.190	56	44	0
A14	0101022 / 2	0.399	0.294	88	12	0
A15	1211112 / 2	0.743	0.610	88	12	0
A16	2111111 / 1	0.189	0.198	60	40	-1
A17	2012120 / 0	0.153	0.230	36	64	0
A18	0000000 / 1	0.679	0.488	88	66	0
A19	1222222 / 2	0.911	0.887	48	52	1
A20	1001200 / 1	0.641	0.719	100	0	-1
A21	1020000 / 0	0.972	0.888	72	28	0
A22	2220121 / 2	0.145	0.042	72	28	0
A23	1000101 / 1	0.952	0.842	72	28	0
A24	2212222 / 2	0.144	0.117	56	44	0
A25	2220020 / 0	0.283	0.329	44	56	0

Table 4-2: Evaluation of the shots presented in Batch B

CLIP	SEQUENCE	BEAM SCORE	SING. SCORE	BEAM VOTE	SING. VOTE	DIFF
B01	2102000 / 2	0.582	0.435	56	44	0
B02	1100111 / 1	0.867	0.861	60	40	0
B03	2110100 / 2	0.998	0.985	96	4	0
B04	0120200 / 0	0.020	0.009	32	68	1
B05	0110000 / 0	0.252	0.291	60	40	-1
B06	1110201 / 1	0.996	0.993	44	56	1
B07	1112221 / 1	0.275	0.218	56	44	0
B08	0111110 / 1	0.816	0.693	64	36	0
B09	0101011 / 0	0.045	0.065	44	56	0
B10	1111011 / 1	0.041	0.059	64	36	-1
B11	1001111 / 0	0.649	0.587	20	80	1
B12	1102022 / 0	0.365	0.373	44	56	0
B13	2111111 / 1	0.712	0.659	68	32	0
B14	1111011 / 2	0.616	0.389	16	84	1
B15	1112021 / 1	0.398	0.430	80	20	-1
B16	2012212 / 2	0.190	0.288	92	8	-1
B17	1200112 / 2	0.592	0.543	56	44	0
B18	1212112 / 0	0.952	0.927	8	92	1
B19	2222222 / 1	0.690	0.097	24	76	1
B20	1120111 / 2	0.972	0.887	8	92	1
B21	0122000 / 0	0.182	0.079	40	60	1
B22	1111100 / 1	0.283	0.122	72	28	0
B23	0111100 / 1	0.511	0.544	80	20	-1
B24	1120000 / 1	0.173	0.006	84	16	0
B25	1200100 / 0	0.953	0.967	32	68	0

4.3 Significance Test

The transformer's results and the user test's results were used to understand whether both parts think alike or if the choices of video variants are significantly different.

4.3.1 Model-User Comparison

The model and the user test answered two different questions. While the model gave a score to each variant independently ("what do you think about this variant?"), the users had to choose between two variants ("which variant of the two is better?"). To make the results comparable, we check whether both the model and the users gave a higher score and thus preferred the same variants and calculate the difference. If both chose the same variant, the difference is 0. If the model chose the beam version and the users were in favor of the singular, the difference is 1. The opposite, model choosing singular and users beam, is -1.

4.3.2 Paired Samples t-Test

We conduct a two-tailed paired samples t-test to check whether the mean value of two dependent samples differ significantly from each other, to compare the mean of a single group (the 50 clips) examined in two different times [23], by the model and by the users. We examine our hypothesis H_0 and the alternative hypothesis H_1:

- H_0: There is no relation between the decisions of the model and the users. Any agreements or disagreements are due to coincidence.
- H_1: There is a relation between the decisions of the model and the users. The model and the users make the same decisions (or opposite decisions).

The test is conducted as follows:

Significance Level	$\alpha = 0.05$
Number of Cases	$n = 50$
Degrees of Freedom	$df = 49$
Mean of Differences	$\bar{x} = -0.1$
Standard Deviation	$s = 0.6776$
Standard Error of The Mean	$s_e = 0.0958$
t-value	$t = 1.0435$
p	$p = p(x \leq 1.0435) = 0.8491$
p-value	$p_v = 2 * \min(p, 1 - p) = 2 * \min(0.8491, 0.1509) = 0.3018$

Since p-value > α, H_0 cannot be rejected. Agreeing on 26 out of the 50 clips, the test concludes that a relation between the decision of the model and the users cannot be proved, and it is likely that data about more clips and votes from more users are required. Looking closer and comparing the model's decisions to the users' decisions per clip, we can find interesting aspects about what humans care about that the model misses. This will be presented in the next chapter.

5 Discussion and Analysis

In this chapter, we discuss and analyze the results, with focus on the similarities and differences between the decisions of the model and the users', by looking on the groups as defined at section 4.2: large majority, medium majority and small majority, with specific clips as examples. For easier comparison, each chosen clip will be presented by its retargeting beam sequence, the singular operator, the model's scores and the users' votes. In each group, both cases where the model and the users agreed with each other as well as cases where the two disagreed will be discussed.

The average lengths of the clips in the large, medium and small groups were 115, 101 and 103 frames respectively. There is therefore no obvious relation between clip length and majority vote group.

It is important to remember that the model, based of its knowledge and circumstances, such as the spatial and spatiotemporal complexity of the clip or the decided hyperparameters, tries to present the best possible retargeted version of the clip. However, it might be impossible to obtain a prefect retargeted version of a clip. While the user's chose the best version out of two possibilities, it is possible that none of those fulfill or are close to fulfill the requirements from a retargeted video, as described in the problem definition, and in some cases the model has also declared the most optimal solution as suboptimal, with score smaller than 0.5.

5.1 Large Majority

A total of 20 clips were classified as large majority. Out of them, 10 had an agreement between the model and the users (50%). When singular got a large majority vote from the users', it was never the choice of the model. Table 5-1 presents the selected clips that will be discussed in this section, while Figure 5-1 presents screenshots of the clips and variants.

Table 5-1: **Selected clips with large majority together with their beam sequence, model scores (probabilities) and user votes (%)**

	SEQUENCE	BEAM SCORE	SINGULAR SCORE	BEAM VOTE	SINGULAR VOTE
B03	2110100 / 2	**0.998**	0.985	**96**	4
A18	0000000 / 1	**0.679**	0.488	**88**	12
B24	1120000 / 1	**0.173**	0.006	**84**	16
A21	1020000 / 0	**0.972**	0.888	**72**	28
A20	1001200 / 1	0.641	**0.719**	**100**	0
B18	1212112 / 0	**0.952**	0.927	8	**92**
A12	1011102 / 1	0.754	**0.767**	**84**	16
B19	2222222 / 1	**0.690**	0.097	24	**76**

Figure 5-1: **Screenshots of the selected large majority clips (upper: beam variant, lower: singular variant). The variant chosen by the users is framed. If the model chose the same, the frame is green, otherwise it is red.**

Large majority clips, that were the easiest for users to decide on, often have a main region of interest and feature humans, in particular faces, or text as the main motives. As mentioned earlier, change of less than 5% in the aspect ratio of faces is enough to result in unnatural appearance. Furthermore, users who were familiar with the actors that were featured in some of the clips, for example Jonah Hill (A12) could easily identify that his appearance was unnatural, and the beam variant, which is less scaled than the singular variant, was preferred. This is an interesting point to remember in case the audience is familiar with the material to be retargeted.

Further examples where a close-up face was distorted and users preferred the most natural looking version (A18, B24), even in case of an animated film where characters' look does not have to agree with real life. Unlike A12, where the model preferred the singular version by a tiny margin, in these two had an agreement.

Distortions in clips featuring groups of people were also noticeable, with the users preferring the beam variant of A20 compared to the scaled singular variant. In B18, the beam variant did not only scale the frame but also used seam carving which caused artifacts that can be seen on the man standing in the middle. In that case, users preferred the cropped version, which features less information, but is not distorted. Artifacts caused by seam carving can be also seen on the fruit and board of B03, while the couple in the frame looks the same in both variants. The decision for the users was obvious.

Text is the main motive of both A21 and B19. In A21, the model could understand that crop for the text is no good, but in B19 the model's choice, which was the beam result, was complete seam

carving, that seem unnatural. This was however compared to scaling as the singular variant, which preserves the content. The result might have been different if it was seam vs crop then users would prefer skewed text to incomplete.

5.2 Medium Majority

A total of 17 clips were classified as medium majority. Out of them, the model and the users agreed on seven. Medium majority clips usually have multiple ROIs and show artifacts and distortions that are less obvious to the viewer or appear in less important areas of the frame. As these clips feature multiple ROIs, it might not be obvious to spot the differences while inspecting short clips. Table 5-2 presents the selected clips that will be discussed in this section, while Figure 5-2 presents screenshots of the clips and variants.

Table 5-2: **Selected clips with medium majority together with their beam sequence, model scores (probabilities) and user votes (%)**

	SEQUENCE	BEAM SCORE	SINGULAR SCORE	BEAM VOTE	SINGULAR VOTE
A02	0011202 / 1	**0.594**	0.512	**68**	32
B08	0111110 / 1	**0.816**	0.693	**64**	36
A17	2012120 / 0	0.153	**0.230**	36	**64**
A06	1110000 / 0	**0.981**	0.975	40	**60**
A03	0222101 / 1	**0.799**	0.638	32	**68**
B04	0120200 / 0	**0.020**	0.009	32	**68**

Figure 5-2: **Screenshots of the selected medium majority clips (upper: beam variant, lower: singular variant) The variant chosen by the users is framed. If the model chose the same, the frame is green, otherwise it is red.**

Clips A17 and A06 both showed a cropped singular variant against a beam version that included also other retargeting operators. In both cases, the users preferred the singular version to the beam, but the medium majority showed that less users could notice that.

Clips B08 featured no humans but a landscape with text. The beam variant which featured one crop at the beginning, one crop at the end and five scales, seemed a much better alternative to only scale, agreed by both the model and the users. Even though less landscape is visible because of the crops, it was the location and size of the text that mattered. Unlike the large majority clips A21 and B19, no seam carving was involved in these variants which avoided the insertion of (obvious) artifacts in the background and/or the text.

Clips A02, A03 and B04 had a beam variant with artifacts in minor ROIs caused by seam carving. In A02, the decision had to be made between a distorted doorway (beam variant) or a scaled face (singular variant). Both model and users preferred the beam variant in this case. In A03, the distorted pattern on the house made the users choose a scaled singular variant, while in B04, users preferred a full crop to a couple of seam carvings that need a closer look to notice (on the man to the right).

5.3 Small Majority

A total of 13 clips were classified as small majority clips. The model and the users agreed on the variants of ten of them. As expected, small majority clips are characterized by unnoticeable technical or visual differences between the variants, quick motion, or artifacts in minor ROIs, which made it more difficult for the users to have a clear preference. In some cases, the users' decisions were made more or less randomly or by gut feeling, which can result in an expected outcome of close to 50-50 in the user vote. Table 5-3 presents the selected clips that will be discussed in this section, while Figure 5-3 presents screenshots of the clips and variants.

Table 5-3: **Selected clips with small majority together with their beam sequence, model scores (probabilities) and user votes (%)**

	SEQUENCE	BEAM SCORE	SINGULAR SCORE	BEAM VOTE	SINGULAR VOTE
A24	2212222 / 2	**0.144**	0.117	**56**	44
B07	1112221 / 1	**0.275**	0.218	**56**	44
A04	2121121 / 1	**0.670**	0.656	**52**	48
B12	1102022 / 0	0.365	**0.373**	44	**56**
A07	1110121 / 1	**0.514**	0.471	48	**52**
A19	1222222 / 2	**0.911**	0.887	48	**52**

Figure 5-3: **Screenshots of the selected small majority clips (upper: beam variant, lower: singular variant) The variant chosen by the users is framed. If the model chose the same, the frame is green, otherwise it is red.**

Clips A24 and A19 both underwent six seam carving operations in the beam version and one scaling operation. Compared to the singular version, which was complete seam carving, the model preferred the variant that was less scaled. While users could identify more easier the variant which was less scaled in A24 (clear ROI), this was more difficult in A19.

Clips B07 and B12 featured high motion scenes. In B07, both the model and the users preferred a version that was a bit more distorted by seam carving, although hardly noticeable, to a singular scale, which changed the proportions of the character. The distortion of the house to the left in B12, caused by seam carving, along with the changed proportions of the humans by scaling, were quite more noticeable but still non-trivial due to the high motion of the clip, and the preferred variant by both was the singular crop.

In A04, the beam variant included a decent amount of scale, compared to a scaled singular variant. As seam carving affected a minor ROI in the clip, the middle shadow, and not the characters, the preference was not significant. The differences in A07 are even more unnoticeable as the clip presents artistic graphics with a large black area.

5.4 Model-User Decision Making

The majority groups, as explained in section 4.2, are divided according to the scores given to the preferred variant by the users. Looking at the model's scores in relation to these groups, it is possible to find patterns and similarities between the user and model decisions.

Calculating the average user score for the chosen variants in each group, it is obvious that the large group will have a higher average than medium and small, but surprisingly, the averages of the model scores for the chosen variants in each group follow the same pattern (see Table 5-4), even if the actual choices were not always the same.

Table 5-4: **Min, Max and Average values of the votes given by the users and the scores given by the model for their preferred variant of each clip, according to the majority groups**

	USERS			MODEL		
	MIN	MAX	AVG	MIN	MAX	AVG
LARGE	72	100	**85**	0.145 (A22)	0.998 (B03)	**0.624**
MEDIUM	60	68	**64**	0.020 (B04)	0.999 (A01)	**0.520**
SMALL	52	56	**55**	0.020 (A05)	0.996 (B06)	**0.436**

Another feature that can be compared is the determination of the choice between the two variants of every clip, by calculating the absolute value of the difference between the scores of every clip's two variants. As the clips were divided to groups from the beginning according to this feature, it is obvious that the average difference of the users will be higher at the large group. However, the model, which gave scores to the variants independently, could more easily discriminate between the variants of the large group, having the highest average score difference for the clips at the large group and smallest average score for the small group, even if the actual choices were not always the same.

Table 5-5: **Min, Max and Average values of the vote differences (users) and score differences (model) between variants of each clip**

	USERS			MODEL		
	MIN	MAX	AVG	MIN	MAX	AVG
LARGE	44	100	**69**	0.013 (B03)	0.592 (B19)	**0.117**
MEDIUM	20	36	**28**	2.145×10^{-6} (A01)	0.160 (A03)	**0.051**
SMALL	4	12	**10**	0.003 (B06)	0.146 (B01)	**0.035**

6 Conclusions and Future Work

This thesis project examined whether a deep learning model, in this case a vision transformer, can understand how to apply a sequence of retargeting operators to a video in a way that the outcome will seem as natural as possible to the viewer, in relation to the definition as presented in section 1.2. A model that can understand what users will care about while watching a video can therefore assist in retargeting videos in a way that will answer the problem definition as good as possible, in relation to the spatiotemporal complexity of the content.

6.1 Conclusions

The beam variant got a majority of the user votes in 30 out of the 50 tested clips, while the model preferred 35 beam variants, which means that both agree that a combination of operators, was most of the time, a better solution. Out of the 20 clips where users preferred the singular variant, 11 were cropped out of 14 presented crops (78.5%), 6 were scaled out of 24 presented scales (25%) and only 3 were seam carved, out of 12 seam carvings (25%). The wide preference of crop as a singular variant makes sense as it does not distort the structure of the video nor create artifacts.

The results show that human perception is very sensitive to scaling, especially when it affects major ROIs, faces, humans and text. Seam carving is more noticeable in major ROIs and especially when straight lines are distorted, as presented in Figure 2-3. The ability to have an unambiguous decision between the variants also depended on the spatiotemporal complexity of the video, e.g., the amount of action, motion and ROIs.

In 23 of the clips the scaled singular variant was presented, which means that the model decided that it was a better option than crop or seam carving. Knowing that humans are more sensitive to scale and tend to prefer crop, is possible that the users' votes could be different if more cropped singular versions were presented. A study examining deepfake videos in five difficulty categories showed that humans and machines have different ways of detecting deepfakes so that videos that are clearly fake for humans were not detected by the machine and vice versa, concluding that human vision and machine vision are different and do not correlate [24]. This can explain a part of the disagreements between the model and the users, that operations which resulted in unnoticeable changes for humans have been important for the model and vice versa.

The same model was trained and tested on a big variety of content. It might be useful to train and obtain specific weights for specific retargeting purposes, depending on the content or the desired aspect ratio change. As smaller data amounts are needed for fine-tuning transformers, it might be feasible to create special training datasets for various retargeting tasks.

Despite the differences and that the model did not manage to develop a human-like vision and perception in this case, the ability to analyze videos and make complex decisions can be in favor deep learning for video retargeting purposes in the future.

6.2 Limitations

This study used both the retargeting operators and the transformer as-is, and they were treated as black boxes. It is possible that optimizing the operators, for example using a better version of seam carving or tweaking the transformer architecture could provide better results.

Furthermore, only shots were examined, but the main purpose of video retargeting is to retarget complete videos, e.g., tv-programs or films. Thus, retargeting longer clips that combine consecutive shots can provide an understanding about video retargeting which is closer to the ideal real-life scenarios.

Collecting more data, both in terms of clips and in terms of users, can be helpful to get to more precise insights and conclusions.

6.3 Future Work

The baseline of this thesis and the assumptions that deep learning can help in video retargeting using various operators in a greedy search approach were shown as a good beginning point. The method can be further examined in the following ways:

- Test the same algorithm as described in section 3.3.1 with other hyperparameters and other aspect ratios.
- Fine-tune the model using other tailored data to retarget a specific type of video.
- Use ROI detection or segmentation of the video to layers to understand which areas of the video should not be altered and apply the retargeting on the background or less important areas.
- Use computer vision to detect the content of the video (people, nature, buildings) and apply a suitable retargeting procedure accordingly.

6.4 Reflections

The procedure of video processing alone can be very energy-consuming. Seam carving and seam adding are operations that require a significant number of mathematical operations. Adding the GPU-based transformer for prediction and using a search algorithm where many partial states are being thrown away, it makes the algorithm used in the thesis non-optimal energy wise for large scale video retargeting, both economically and environmentally. As seen in the retargeting sequences of the clips that were a part of the test, it is possible that the algorithm will come to a beam search output that is in fact a singular operation or something that is very close to it (for example if 6 out of 7 operations were the same operator).

A suggestion for future development is to do a pre-analysis of the material and figure out which scenes can be retargeted using a singular operation and which more complex scenes require such computational heavy retargeting.

Although it was mentioned that seam carving can be used to create fake but realistic representations of images, in videos it might be more difficult, based on the limited success of seam carving as a singular operator in this study. It is hard to believe that video retargeting will be used to create deepfake videos, as there are other areas of deep learning which specializes in this, but it is not unrealistic that other concepts from video retargeting, with or without deep learning, will be used in the future for other purposes.

References

[1] Kopf, S., Haenselmann, T., Kiess, J., Guthier, B., & Effelsberg, W. (2011). Algorithms for video retargeting. Multimedia Tools and Applications, 51(2), 819-861. DOI: http://doi.org/10.1007/s11042-010-0717-6

[2] Jack, K. (2011). Video demystified: a handbook for the digital engineer (5th ed.). Elsevier. 223-232. ISBN: 978-0-7506-8395-1.

[3] Vaquero, D., Turk, M., Pulli, K., Tico, M., & Gelfand, N. (2010, September). A survey of image retargeting techniques. In Applications of Digital Image Processing XXXIII (Vol. 7798, p. 779814). International Society for Optics and Photonics. DOI: http://doi.org/10.1117/12.862419

[4] Avidan, S., & Shamir, A. (2007). Seam carving for content-aware image resizing. In ACM SIGGRAPH 2007 papers (pp. 10-es). DOI: https://doi.org/10.1145/1275808.1276390

[5] Rubinstein, M., Shamir, A., & Avidan, S. (2008). Improved seam carving for video retargeting. ACM transactions on graphics (TOG), 27(3), 1-9. DOI: https://doi.org/10.1145/1360612.1360615

[6] Dong, W., Zhou, N., Paul, J. C., & Zhang, X. (2009). Optimized image resizing using seam carving and scaling. ACM Transactions on Graphics (TOG), 28(5), 1-10. DOI: https://doi.org/10.1145/1618452.1618471

[7] Song, E., Lee, M., & Lee, S. (2018). CarvingNet: content-guided seam carving using deep convolution neural network. IEEE Access, 7, 284-292. DOI: https://doi.org/10.1109/ACCESS.2018.2885347

[8] Shamir, A., & Sorkine, O. (2009). Visual media retargeting. In ACM SIGGRAPH ASIA 2009 Courses (pp. 1-13). DOI: https://doi.org/10.1145/1665817.1665828

[9] Tekalp, A. M. (1995). Digital video processing. Prentice-Hall, Inc.. 348-349. ISBN: 978-0-1319-0075-2

[10] Lin, H., Si, J., & Abousleman, G. P. (2007, April). Region-of-interest detection and its application to image segmentation and compression. In 2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems (pp. 306-311). IEEE. DOI: http://doi.org/10.1109/KIMAS.2007.369827

[11] Liu, H., Jiang, S., Huang, Q., Xu, C., & Gao, W. (2007, September). Region-based visual attention analysis with its application in image browsing on small displays. In Proceedings of the 15th ACM international conference on Multimedia (pp. 305-308). DOI: https://doi.org/10.1145/1291233.1291298

[12] Rubinstein, M., Shamir, A., & Avidan, S. (2009). Multi-operator media retargeting. ACM Transactions on graphics (TOG), 28(3), 1-11. DOI: https://doi.org/10.1145/1531326.1531329

[13] Gonzalez, R. C., & Woords R. E. (2018). Digital image processing (4th ed.). Pearson. 77-78. ISBN: 978-0-1333-56724-1.

[14] Westerlund, M. (2019). The emergence of deepfake technology: A review. Technology Innovation Management Review, 9(11). https://timreview.ca/article/1282

[15] MasterClass. Guide to Aspect Ratios: 8 Film and TV Aspect Ratios. 21 June 2021. Retrieved 4 May 2022. https://www.masterclass.com/articles/guide-to-aspect-ratios#what-is-an-aspect-ratio-in-film

[16] Kelleher, J. D. (2019). Deep learning. MIT press. ISBN: 978-0-2625-3755-1.

[17] Zhu, C. (2019). Fast video retargeting based on seam carving with parental labeling. arXiv preprint. arXiv: https://arxiv.org/abs/1903.03180

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://dl.acm.org/doi/10.5555/3295222.3295349

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929

[20] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022). https://doi.org/10.48550/arXiv.2103.14030

[21] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2021). Video Swin Transformer. https://doi.org/10.48550/arXiv.2106.13230

[22] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. https://doi.org/10.48550/arXiv.1212.0402

[23] Ross, A., & Willson, V. L. (2017). Paired samples T-test. In Basic and advanced statistical tests (pp. 17-19). SensePublishers, Rotterdam. https://doi.org/10.1007/978-94-6351-086-8_4

[24] Korshunov, P., & Marcel, S. (2020). Deepfake detection: humans vs. machines. arXiv preprint. arXiv: https://arxiv.org/abs/2009.03155

For DIVA

{”Author1”: { ”Last name”: ”Laufer”,
”First name”: Gil”,
”Local User Id”: ”u19cbacn1”,
”E-mail”:”glaufer@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,}},

”Cycle”: ”2”,

”Course code”: ”DA233X”,

”Credits”: ”30.0”,

”Degree1”: {”Degree”: ”Master’s degree”,
”Educational program”: ”Master’s Programme in Machine Learning”,
”programcode”: ”TMAIM”,
”subjectArea”: ”Computer Science and Engineering”},

”Title”: {”Main title”: ”This is the title in the language of the thesis”,
”Subtitle”: ”Utilizing deep learning for video aspect ratio change”,
”Language”: ”eng” },

”Alternative title”: {”Main title”: ”Video Retargeting med hjälp av Vision Transformers”,
”Subtitle”: ”Användning av djupinlärning för ändring av videobildförhållanden”,
”Language”: ”swe”},

”Supervisor1”: { ”Last name”: ”Beskow”,
”First name”: ”Jonas”,
”Local User Id”: ””,
”E-mail”: ”beskow@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”}},

”Examiner1”: { ”Last name”: ”Sturm”,
”First name”: ”Bobby Lee Townsend”,
”Local User Id”: ””,
”E-mail”: ”bobs@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”}},

”Cooperation”: { ”Partner_name”: ” Entecon AB”},

”Other information”: {”Year”: ”2022”, ”Number of pages”: ”xi,39”},

”Series”: { ”Title of series”: ”TRITA-EECS-EX” , ”No. in series”: ”20XX:XX” },

”Number of lang instances”: "2",

”Abstract[eng]”: €€€€

The diversity of video material, where a video is shot and produced using a single aspect ratio, and the variety of devices that can play video with screens in different aspect ratios make video retargeting a relevant topic. The process of fitting a video filmed in one aspect ratio to a screen in another aspect ratio is called video retargeting, and the retargeted video should ideally preserve the important content and structure of the original video as well as be free of visual artifacts. Important content and important structure are vague and subjective definitions, which makes this problem more difficult to solve. The video retargeting problem has been a challenge for researchers from the computer vision, computer graphics and human-computer interaction areas, and successful retargeting can improve the viewing experience and the content’s aesthetic value. Video retargeting is done by four tools: cropping, scaling, seam carving and seam adding. Previous research showed that one of the keys to successful retargeting is to use a suitable combination of operators. This study makes use of a vision transformer, a deep learning model which is trained to discriminate between original and retargeted videos. Solving an optimization problem using beam search, the transformer assists in choosing a combination of operators that will result in the best possible retargeted video. The retargeted videos were examined in a user A/B-test, where users had to choose their preferred variant of a video shot: the transformer’s output using beam search, or a singular version where the video underwent a single retargeting operation. The model and user preferences were compared to check if the model indeed can make retargeting decisions that are appealing for humans to watch. A significance test showed that no conclusion can be made, probably due to lack of enough test data. However, the study revealed patterns in the preferences of the users and the model that could be further fine-tuned or combined with other computer vision mechanisms in order to output better retargeted videos.

€€€€,

”Keywords[eng]”: €€€€

Video retargeting, Aspect ratio, Computer vision, Deep learning, Vision transformers.

€€€€,

”Abstract[swe]”: €€€€

Variation av videomaterial, där olika videor är inspelade och producerade i olika bildförhållande, samt variation i apparater och skärmar som spelar upp videor i olika bildförhållanden gör ändring av videobildförhållande till en relevant fråga. Processen där en videos bildförhållande ändras heter video retargeting. När video retargeting används bör den nya videon helst bevara strukturen och viktigt innehåll från originalvideon samt vara artefaktfri. Struktur och viktigt innehåll är subjektiva definitioner vilket gör frågan svårlöst, och frågan har varit en utmaning för forskare inom datorseende, datorgrafik och människa-datorinteraktion. Lyckad ändring av en videos bildförhållande kan förbättra tittarupplevelsen och innehållets estetiska värde. Video retargeting kan göras med hjälp av fyra funktioner: klippning, skalning, seam carving och seam adding. Tidigare studier visar att en av nycklarna till lyckad retargeting är att hitta en lämplig kombination av funktionerna. I denna studie används Vision Transformer, en djupinlärningsmodell som tränas för att skilja mellan original och omvandlade videor. Genom att lösa ett optimeringsproblem med strålsökning hjälper modellen välja den kombination av funktionerna som resulterar i den bästa möjliga omvandlade videon. De omvandlade videorna testades genom ett användartest där användare valde vilket videoklipp de tyckte bättre om: modellens output som skapades med hjälp av strålsökning, eller en version där klippet genomgick en enklare retargeting med hjälp av endast en av funktionerna. Modellens och användarnas preferenser jämfördes för att se om modellen kan fatta beslut som användare upplever som bra. Ett signifikanstest visar att ingen slutsats kan dras, förmodligen på grund av det begränsade antalet videoklipp och data som användes i studien. Däremot visar studien mönster i användarnas och modellens preferenser som kan användas för att vidareutveckla problemlösningen inom området.€€€€,

”Keywords[swe]”: €€€€

Video retargeting, Bildförhållande, Datorseende, Djupinlärning, Vision transformers.

€€€€,

}