Automated Objective Video Quality Assessment for the Automated Edinburgh Visual Gait Score (EVGS)

Rajkumar Arumugam Jeeva

Thesis submitted to the University of Ottawa in partial fulfillment of requirements for the degree of

Master of Applied Science

in

Biomedical Engineering

uOttawa

Ottawa Carleton Institute for Biomedical Engineering
University of Ottawa
Ottawa, Ontario

Abstract

This thesis presents a framework for automated video quality assessment in gait analysis, addressing critical challenges in clinical gait evaluation. The research focuses on developing robust algorithms for detecting multiple persons, plane orientation, zoom artifacts, computing overall video quality score, and generating feedback suitable for automated Edinburgh Visual Gait Score (EVGS) scoring.

The methodology involves extracting skeletal keypoints from video frames using the MoveNet Lightning pose estimation model. The algorithms use these keypoints to detect multiple persons, track the person of interest, detect the plane of motion, identify overlapping people, detect camera zooming, and evaluate video resolution. These components are integrated into a unified quality classification system using a Random Forest classifier.

Results demonstrated exceptional performance across various metrics. The plane detection algorithm achieved excellent classification in both training and validation datasets, ensuring only appropriately videos are used for analysis. Multiple person detection and overlap assessment algorithms show strong performance, with accuracies of 96% and 95% respectively in the validation dataset. The zoom detection algorithm achieved 92% accuracy in identifying sudden zoom events in the video. The overall video quality assessment framework demonstrates a 95% accuracy in categorizing videos as either suitable for immediate analysis or requiring manual editing. This high level of accuracy showcases the effectiveness of the proposed methodology in automating the quality assessment process. The system also provides specific suggestions for improvement when videos fail to meet quality standards, enhancing the overall efficiency of the gait analysis workflow.

In conclusion, this research makes significant contributions to automated gait analysis by addressing the crucial aspect of video quality assessment. The developed framework demonstrates high performance score across various quality metrics, potentially removing the need for manual quality check for automated EVGS scoring in clinical practice

Table of Contents

List of Figures

List of Tables

Abbreviations and Definitions

CHO	Channelized Hotelling Observer
CNN	Convolutional Neural Network
CT	Computed Tomography
DSCQS	Double-Stimulus Continuous-Quality Scale
EVGS	Edinburgh Visual Gait Score
FN	False Negatives
FP	False Positives
FR	Full Reference
HSD	Horizontal Shoulder Distance
HVS	Human Visual System
LVSD	Left Vertical Shoulder-to-Hip Distance
MCC	Matthews Correlation Coefficient
MOS	Mean Opinion Score
MSE	Mean Squared Error
NPWMF	Non Pre-Whitening Matched Filter
NR	No Reference
PSNR	Peak Signal-to-Noise Ratio
ROC	Receiver Operating Characteristics
RR	Reduced Reference
RVSD	Right Vertical Shoulder-to-Hip Distance
SNR	Signal to Noise Ratio
SSIM	Structural Similarity Index
TN	True Negatives
TP	True Positives
VIF	Visual Information Fidelity
VQA	Video Quality Assessment
VQM	Video Quality Metrics
VSNR	Visual Signal to Noise Ratio

Acknowledgments

I sincerely thank and appreciate my supervisors, Prof. Natalie Baddour and Prof. Edward Lemaire, for their unwavering support and guidance throughout my research journey. Their belief in me and the opportunity to work at the Mobile Motion Lab were instrumental in the completion of this thesis. Their weekly support, even during slow progress periods, along with their informative guidance and patience, were invaluable. They not only assisted me in becoming more organized but also taught me to think like a researcher when approaching problems.

I extend my heartfelt gratitude to Dr. Albert Tu and Dr. Kevin Cheung from CHEO, whose vast clinical knowledge and expertise in EVGS were crucial to this research. Their continuous assistance and contributions were indispensable to the success of this project.

The greatest gratitude goes to my family, especially my parents, Arumugam Jeeva, and Nirmala Jeeva, my brother Vimalkumar Arumugam Jeeva, and my sister-in-law Vishali Vimalkumar. Their help, guidance, and, most importantly, their unconditional love has always inspired me to express myself more clearly and accomplish all my goals.

Lastly, I extend my sincere thanks to all my friends and colleagues at The University of Ottawa, Ottawa Hospital Rehabilitation Centre, The Children's Hospital of Eastern Ontario and doctors of Sanatorio del Norte for their invaluable assistance with data collection and validation. Your support and collaboration have significantly contributed to the success of this research.

1 Introduction

1.1 Research context and scope

Gait analysis has emerged as a crucial tool in clinical settings for diagnosing and treating various gait-related pathologies [1]. The systematic study of human locomotion provides invaluable insights into biomechanical stability and motor function, supporting surgical planning and rehabilitation therapies [2]. While traditional gait analysis methods have relied on manual observation and scoring, recent advancements have led to automated systems for recognizing gait events and estimating gait parameters [3]. However, the effectiveness of these automated systems heavily depends on the quality of input videos.

Video quality assessment (VQA) has become increasingly important in the context of medical imaging and gait analysis [4]. Unlike entertainment or general-purpose videos, medical videos require specialized quality metrics that consider diagnostic relevance and clinical utility. In the realm of gait analysis, video quality can significantly affect the accuracy of automated assessments, making it crucial to develop robust VQA methods tailored to this domain [4], [5]. Content-based video quality assessment has evolved to incorporate both technical and semantic elements, recognizing that quality extends beyond mere technical parameters [6]. The presence of artifacts like zoom effects, more than one person, or invalid camera angle can severely affect automated gait analysis algorithm performance.

Recent research has focused on developing task-specific quality assessment methods for medical videos [6]. These approaches aim to evaluate video quality based on its fitness for specific clinical tasks, such as gait parameter estimation or joint angle measurement. By incorporating domain knowledge and clinical requirements, these methods provide more relevant quality scores than traditional, general-purpose VQA techniques [4], [5], [7].

This thesis encompasses the development and validation of a comprehensive video quality assessment framework specifically designed for gait analysis videos. This framework addresses key challenges in automated gait analysis, including the detection of multiple persons, plane orientation classification, zoom artifact identification, and resolution assessment. By providing an objective measure of video quality, and generating feedback to users, this research seeks to eliminate the need the manual quality checking, making automated EVGS scoring truly automatic.

1.2 Rationale

Automated systems for recognizing gait events and estimating EVGS scores represents a substantial advancement in the field [3]. These systems leverage pose estimation models, such as OpenPose [8], to extract body keypoints and assess gait patterns. However, the effectiveness of these automated systems heavily depends on the quality of input videos [9]. Unlike manual scoring, where human observers can adapt to variations in video quality, automated systems are limited in their ability to cope with poor input conditions [10]. Factors such as bad camera angles, zooming artifacts, low resolution, and the presence of multiple individuals can adversely effect keypoint detection accuracy. Consequently, high-quality input videos are essential for reliable automated EVGS scoring. The need for consistent, high-quality video inputs has led to the current practice of manual quality checks, which are time-consuming, subjective, and prone to inconsistencies. Automating the video quality assessment process is crucial to improve efficiency, reduce human bias, and ensure consistent and accurate inputs for automated EVGS systems.

1.3 Scope

This thesis focuses on developing and validating an automated video quality assessment framework for gait analysis videos. The research encompasses the design, implementation, and validation of algorithms for detecting multiple persons, plane orientation, zoom artifacts, and overall video quality suitable for automated EVGS scoring.

1.4 Objective

The primary objective is to create a robust, automated system for assessing the quality of gait analysis videos.

1.5 Thesis contributions

This thesis makes several important contributions to the field of automated gait analysis and video quality assessment:

- Developed a comprehensive automated video quality assessment framework specifically tailored for gait analysis videos. This framework integrates multiple quality metrics, including plane detection, multiple person detection, zoom artifact identification, and resolution assessment.
- Created a novel plane detection algorithm that accurately classifies video orientations into

sagittal, coronal, and transverse planes. Defining the plane is important since the automated EVGS algorithm applies different measurement techniques specific to coronal and sagittal planes.

- Created a robust multiple person detection and overlap assessment algorithm, addressing the challenge of identifying and isolating the person of interest in gait videos.
- Designed an innovative zoom detection algorithm that identifies and quantifies zoom artifacts, a critical factor in maintaining consistent participant size and position for accurate gait analysis.
- Integrated the individual quality components into a unified quality scoring system, that classifies videos as either suitable for immediate analysis or requiring manual editing or redoing.
- Developed an automated feedback mechanism that provides actionable suggestions for improving video quality, enhancing the overall efficiency of the gait analysis workflow.

These contributions collectively address the critical need for automated quality assessment in gait analysis, potentially revolutionizing the automated EVGS scoring in clinical practice.

2 Literature review

2.1 Gait analysis

Gait analysis is the systematic study of human locomotion, often used in a clinical setting to diagnose and treat gait-related pathologies [1]. Observing the process of walking allows a clinician to identify complications and assess treatment outcomes in various therapies or interventions. This analysis is important for treating gait disorders, cerebral palsy, balance impairments, stroke, Parkinson's disease, and orthopedic disorders [3], [11]. Comprehension of human gait provides necessary knowledge for biomechanical stability and motor function for surgical planning and rehabilitation therapies [12].

2.1.1 Edinburgh Visual Gait Score

The Edinburgh Visual Gait Score is a systematic, standardized instrument to assess gait abnormalities with video documentation [13], [14]. The EVGS offers numerical data on the severity of gait impairments by scoring critical gait parameters. Such scoring systems are adopted widely in research settings and clinical practice, offering health professionals a mode of enhancing communication while observing changes in gait over time [15].

EVGS requires clinicians to review video recordings of a patient's gait, typically in the sagittal or coronal planes, and score these based on predetermined criteria. While this approach provides standardized evaluation, scoring may be variable, mainly because of the observer’s experience and the level of proficiency [16].

Traditional EVGS scoring methodologies depended on manual analysis of gait videos, a process that is time-consuming and subject to errors. In fact, manual EVGS evaluation from video can take approximately 24.7 minutes, thereby posing a time burden to healthcare professionals [3]. Moreover, human observation is subjective, bringing inconsistency and is influenced by the scorer’s experience and proficiency. These challenges have led to research into automated methods to improve the accuracy and efficiency of EVGS scoring [3] by recognizing gait events and applying algorithms to estimate EVGS scores. Markerless pose estimation methods, such as OpenPose Body 25, extract body keypoints that denote joints and limb segments [8]. This model has served as a base for algorithms that effectively identified strides and foot events, thus providing

a sound framework for the automated scoring of EVGS [3].

2.2 Dependency of input video quality

Automated EVGS scoring using pose estimation models for body keypoint detection [3], [8] works well in controlled environments and requires clear, stable, and well-framed videos. Unlike manual scoring, where human observers can adapt to these variations in video quality, automated systems are currently limited by an inability to cope with poorer input conditions. Bad camera angles, zooming artifacts (e.g., inward zoom or outward zoom), or low resolution may adversely effect keypoint detection accuracy; hence, high-quality input videos are required for reliable EVGS scoring [17]

Invalid video not only compromises the precision of gait assessments but also reduces trust in system outputs. Automated EVGS systems require videos to be acquired in sagittal and coronal planes. Videos acquired at other angles, including transverse or oblique views, prohibit the system's ability to correctly calculate EVGS scores. Errors in keypoint recognition result from such missteps. Missed or incorrectly identified keypoints subsequently limit the algorithm's capability to calculate valid gait scores.

Automated EVGS algorithms are based on the correct detection of keypoints, which include shoulders, hips, knees, and ankles [3]. Videos with low resolution, zooming artifacts, overlapping individuals, or excess background clutter may influence algorithm performance and cause misidentification of the person or an incorrect bounding box determination. Misaligned, missing, or distorted keypoints all similarly affect parameters precision (i.e., joint angles, stride length, etc.). Errors in the calculated parameters can lead to incorrect EVGS scores [3].

Manual quality checks can be performed to ensure that videos meet quality requirements. However, this is time-consuming, subjective, and prone to inconsistencies. Automating the manual quality check process will help to improve efficiency, reduce human bias, and guarantee consistent and accurate video inputs for an automated EVGS system, which leads to a truly automated system. Automated video quality assessment mechanisms can be implemented to overcome these challenges and hence open ways for reliable and efficient use of automated EVGS in clinical settings [9].

2.3 Visual based video quality assessment

Video quality assessment methods are essential for evaluating the suitability of videos for various

applications, ranging from entertainment streaming to clinical and forensic analysis. Depending on the availability of an undistorted reference (ideally the original image or video), visual quality-based techniques are typically divided into three categories: Full Reference (FR), No Reference (NR), or Reduced Reference (RR). NR measures evaluate the quality solely using the distorted image. FR metrics forecast the quality by directly comparing the reference and its distorted counterparts. RR metrics compare features representative of the distorted and reference images [4]. Performance is usually evaluated through a statistical comparison with subjective results [18].

2.3.1 Full-Reference methods

Full-Reference methods compare the distorted video with an original, undistorted reference video. These methods typically provide the most accurate quality predictions but are limited in practical applications where the reference is unavailable [4], [19]. Common metrics include:

- Mean Squared Error (MSE): Average squared difference between corresponding pixels in the reference and distorted videos. While computationally simple, MSE does not correlate well with human perception [20].
- Peak Signal-to-Noise Ratio (PSNR): Derived from MSE, PSNR evaluates the ratio of signal power to noise, providing a logarithmic measure of quality. However, PSNR also fails to account for perceptual factors [20].
- Structural Similarity Index (SSIM): Perceptually oriented metric that evaluates luminance, contrast, and structural similarity between frames. SSIM is widely used due to its better alignment with human visual perception [21], [22].

FR metrics have several limitations. They require access to an undistorted reference image, which is often unavailable in real-world scenarios [19]. FR metrics may not always correlate well with human perception, especially for complex distortions [23]. Additionally, FR metrics can be computationally expensive for large datasets or real-time applications. Some FR metrics like PSNR are sensitive to pixel-wise changes but may be poor at capturing structural information [24], [25].

2.3.2 Reduced-Reference methods

Reduced-Reference image and video quality assessment methods for medical applications use partial information from a reference image or video to evaluate the quality of a distorted image or video [4]. The goal is to find a balance between full-reference approaches, which require complete

access to an undistorted reference. RR methods generally extract some features or attributes from the reference image/video and compare the extracted features from the reference image/video with the distorted contents.

One of the most important tasks in RR metric design is to choose the most meaningful and informative features that should be extracted from the reference image or video [26]. The amount of side information required may vary greatly for different RR approaches. There is usually a trade-off between the amount of reference information used and the accuracy in quality prediction. Some common methods for RR quality assessment include statistical models of natural images, extraction of structural information, or perceptually relevant image attributes [26], [27].

RR metrics face challenges in selecting the most relevant features to extract from the reference image. The amount of side information needed can vary between different RR methods. RR metrics may not perform as well as FR metrics for certain types of distortions. Effectiveness can be limited by the quality and relevance of extracted features [28], [29], [30].

2.3.3 No-Reference methods

No-reference quality assessment methods aim to assess the quality of medical images and videos when no access to a reference or undistorted version is possible. NR approaches are especially valuable for real-world applications since reference images are usually not available [4]. NR methods typically rely on statistical models of natural images, machine learning techniques, or analysis of particular artifacts and distortions [31], [32]. Many NR metrics are designed for specific imaging modalities or to identify particular types of artifacts [33]. Recent advances in deep learning have led to a surge in the use of convolutional neural networks (CNNs) and other deep learning architectures for NR quality assessment. These methods automatically learn relevant features for quality prediction directly from the image or video data [33], [34], [35].

NR metrics struggle with the lack of a reference image, making it difficult to assess certain types of distortions accurately [36]. NR metrics often rely on statistical models or machine learning approaches that may not generalize well to new types of distortions or imaging modalities [37], [38]. NR metrics can be sensitive to the training data used and may perform poorly on images with characteristics different from the training set [33], [36], [37]. Additionally, many NR metrics are designed for specific distortion types or imaging modalities, limiting their broad applicability [33].

2.4 Content based quality assessment for videos

Content-based video quality assessment has been applied to various domains, each with unique requirements and challenges. VQA studies [39], [40] often focus on the technical viewpoint, assessing video distortions (i.e., blurring, artifacts) and their effects on quality to compare and direct technological systems like cameras [33], restoration algorithms [41], and compression standards [42]. Clear-textured video should be noticeably higher quality than grainy video (Figure 2.1). However, a number of recent studies found that human quality evaluation of movies was also influenced by preferences for non-technical semantic elements such as composition and content [43]. Human experience with these elements is typically thought of as the aesthetic perspective of quality evaluation [44], which favors the video in Figure 2.1 (b). The content recommendation systems on websites like YouTube or TikTok consider videos with meaningful or highly relevant content as high quality even if the visual clarity of the video is blurry or low resolution[45]. In entertainment videos, viewers are more tolerant of poor technical quality (e.g., low frame rate) if they are highly interested in the content [46].

(a) (b)

Figure 2.1 (a) Clear video with meaningless content, (b) Blurry video with meaningful content [45]

Recent approaches in content-aware VQA have moved beyond traditional low-level features to incorporate semantic content analysis and aesthetic considerations [45]. This shift recognizes that video quality is not solely determined by technical aspects but also by content-related factors. Content-based VQA has been applied to various domains, each with unique requirements and challenges. This section explores several key applications, and the research conducted to address their specific needs.

Medical images/videos are typically large and can be compressed to efficiently manage storage and database. When compressing medical images/videos, it is important to ensure that all the features, like anatomical structures and textures, are retained. Diagnostic image quality is not

solely determined by the image's visual clarity but also by the essential features required for diagnosis. Three quality evaluation metrics categories are objective, subjective, and hybrid [5].

2.4.1 Objective quality measurement

Objective quality measurement provides numerical representations for image quality and a quantitative approach to assessment [47]. Objective metrics defined in Table 2.1 are the most popular metrics used in medical image quality evaluation [5].

One widely used objective metric is the Mean Square Error (MSE), which is the mean of the squared difference between the original image and the processed image. Although easy to calculate, MSE does not correlation well with human perception of image quality [48].

Table 2.1 Quality methods for medical image quality evaluation [5]

Reference	Imaging modality	Quality impairments	Validation	Quality metrics
[50]	MRI	Rician and Gaussian noise, Gaussian blur (5 levels); DCT, JPEG and JPEG2000 compression (5 ratios)	MOS: 4 medical observers (SDSCE)	SNR, PSNR, SSIM, WSNR
[51]	CT	JPEG and JPEG2000 compression (5 ratios)	MOS: 6 radiologists (Double-stimulus DCR)	MSE, local MSE, SNR, SSIM, VSNR, VIF
[52]	Ultrasound videos	H.264 Compression, quantization parameters and packet-loss rates	MOS: 2 medical experts	PSNR, SSIM, VSNR, VIF, WSNR
[53]	Ultrasound videos	HEVC compression (8 quantization levels)	MOS: 4 medical experts and 16 naïve observers (DSCQS)	PSNR, SSIM, UQI, VQM, NQM, VIF, VSNR
[54]	Laparoscopic surgery videos	H.264 compression (4 bit rates)	MOS:9 laparoscopic Surgeons and 16 naïve observers (SSCQE)	VQM, HDR-VDP-2, PSNR
[55]	Endoscopic surgery videos	H.264 compression (11 ratios)	MOS: 14 medical observers (DSCQS)	SSIM,UQI, WSNR,VSNR, HDR-VDP,IFC, MSE, MS-SSIM, PSNR-HVS

Another common metric derived from MSE is Peak Signal to Noise Ratio (PSNR), which expresses the ratio of maximum possible signal power over noise power [49]. PSNR remains in common use because it is simple to calculate but, as with MSE, it has limitations in capturing perceptual quality.

Other metrics, such as the Visual Information Fidelity (VIF) and Visual Signal to Noise Ratio (VSNR) [57], [58] try to incorporate human visual system modeling and may provide more perceptually meaningful quality assessments. A study[5], [59] noted that objective metrics are widely used for quality evaluation of medical images despite their limitations. These metrics provide quick and easy-to-compute quality assessments, making them valuable tools in various medical imaging applications.

A strong criticism of objective quality metrics, as used in the assessment of medical images, is poor correlation with diagnostic quality. These metrics often do not sufficiently meet the requirements for medical image interpretation where diagnostic information should be preserved [5]. This limitation gives rise to metrics that emphasize diagnostic quality over either statistical or perceptual criteria.

2.4.2 Subjective quality assessment

Subjective quality assessment incorporates expert opinions on both perceptual quality and diagnostic information preservation [60], [61]. Three main approaches are [5]:

Mean Opinion Score (MOS): MOS is one of the most straightforward and widely used methods in subjective assessment. Medical experts measure the image perceptual quality and visually analyses the diagnostic information preserved in the image [62]. Sets of images are presented in random order to medical professionals who score each on a 1-5 scale where 1 is poor quality and 5 is best quality. Scores are averaged for statistical comparison [53].

Multiple Reader Multiple case: A number of cases are assessed individually by medical experts of various skill levels. Although comprehensive, this method is subject to assessment inconsistencies and potential biases; therefore, tests must be planned carefully. Various experts with varying skill may view the same image differently [63].

Double stimulus Continuous Quality Scale (DSCQS): DSCQS uses a just noticeable difference method where the experts compare two side-by-side pictures, usually an original and a processed version. The experts provide rankings for both pictures, which are used in calculating

mean scores and other statistics. This method is of value when assessing compressed pictures since DSCQS indicates whether "perceptually lossless" compression has been achieved [64].

The most common criticism of subjective assessment is that the results cannot be reliably reproduced. Different experts may grade the same image with vastly different ratings, having a large effect on the average rating. Even experts making individual ratings for the same image in multiple viewings may have different results [61], [65]. Despite these limitations, subjective assessment remains valuable due to its reliability in evaluating diagnostic quality, which objective metrics often fail to capture effectively.

2.4.3 Hybrid quality assessment

Hybrid quality assessment melds the strengths of objective and subjective assessment techniques [66], to overcome limitations in using either of these methods alone [67], [68]. The process generally includes correlating objective measures with subjective expert evaluations in an effort to determine which measures best correspond to human perception and diagnostic needs [69], [70].

Receiver Operating Characteristics (ROC) Analysis: ROC analysis incorporates the decision threshold of diagnosticians, which determines the relationship between sensitivity and specificity [71]. Sample thresholds are calculated based on expert ratings, usually on a 1-5 scale, which are then used to evaluate feature detection accuracy. ROC analysis is especially effective in lesion detection tasks where higher image quality is associated with an increased probability of lesion detection [72].

Correlation-Based Assessment: This method involves statistically correlating objective quality assessment with the subjective quality assessment. An objective metric that exhibits high correlation with the subjective metric is considered to give high diagnostic quality assessment . Non traditional objective metrics shows better correlation with the subjective metrics than the traditional objective metrics such as peak signal noise ration (PSNR), mean square error (MSE), etc. [73], [74], [75], [76].

Human Visual System (HVS) Integration: This approach integrates human visual system characteristics such as contrast sensitivity, luminance perception, masking effects, and frequency decomposition [77], [78]. HVS models tend to correlate well with human visual performance and diagnostic needs [79].

Limitations of hybrid quality assessment include high computational overhead in processing

big medical image datasets and lack of standard procedures for combining metrics [4]. The approach needs a high investment of time, both in objective measurements and also subjective expert evaluations; hence, this method cannot be used in real-time applications [80]. Further, hybrid quality assessment requires the availability of experts, and varying levels of correlations between different objective metrics with the subjective assessments produce inconsistency in the results [81]. The context-dependent nature of medical images further complicates the establishment of a universal approach, because effectiveness varies across different types of medical images and diagnostic requirements [80].

2.4.4 Task based quality assessment

Task-based quality assessment methods aim to evaluate the quality of medical images and videos by measuring their effectiveness for specific clinical tasks. These methods are designed to approximate the performance of human observers, such as medical experts, rather than predicting mean opinion scores [81]. The underlying paradigm is to quantify the quality of medical content by its effectiveness with respect to its intended purpose [4], [80].

Detection tasks involve a choice between two hypotheses: signal present or signal absent. This is of special importance in low-dose reconstruction methods based on iterative algorithms in tomography [4]. A study [82] applied a Channelized Hotelling Observer (CHO) [83], [84] and an internal noise model to compare detectability indices in low-dose CT images. Five observers using CT phantom images showed that Iterative Model Reconstruction allows for at least 67% dose reduction compared to Filtered Back-projection [82].

In another study [85], quality of low-dose CT scans was assessed using the imQuest software to compare different manufacturers and reconstruction algorithms. A Non-Prewhitening Matched Filter (NPWMF) model observer with an eye filter was used to calculate detectability indices for two detection tasks, one of a large mass in the liver and the other of a small calcification.

More recent work in deep learning has encouraged the development of supervised learning-based approaches to model observer implementation; including, CNN-based denoising strategies on simulated planar scintigraphy images [86], Bayesian Ideal Observer [87], Hotelling Observer [83], [87], CHO, Regularized Hotelling Observer, and NPWMF. High sensitivity and specificity were achieved after training a CNN model in classifying mammography screening image patches as either normal tissue or containing a lesion [88].

Several limitations exist in the detection and classification tasks of the task-based quality assessment. Large datasets with known ground truth are required, especially for medical images, and these datasets can be very time-consuming to develop [89]. The performance of model observers like CHO may not generalize well across different imaging conditions or anatomical variations [90]. Most studies deal with very simple detection tasks and artificial signals that do not mimic complexity in real clinical scenarios. The use of simulated or phantom images allows controlled experimentation but may not correctly represent the statistical properties of real patient data. Deep learning approaches look very promising but usually require large amounts of training data and are less interpretable than traditional model observers [89].

2.4.5 Localization tasks

Localization tasks measure the ability to pinpoint anatomical structures or pathologies from medical images. Approaches include a Localization CNN (L-CNN) to localize the regions of interest for the fetal abdomen, with image quality based on the visibility of some structures [91]; supervised learning to approximate an Ideal Observer in joint detection and localization tasks [92], and a U-Net-based model observer for defect localization on simulated images with different levels of correlated noisy backgrounds [93], where models trained with binary cross-entropy loss functions had results closer to human performance.

Localization tasks in medical image quality assessment are challenged by the definition and precise measurement of localization accuracy. Much of the literature uses simplified localization tasks that do not represent the complexity of clinical practice [94]. Localization tasks are sensitive to anatomical variations between patients, so the generalizability of results obtained on phantom or simulated data may be limited. The development of clinically relevant localization tasks is an active area of research with room for improvement [95], [96].

2.4.6 Segmentation tasks

Segmentation tasks use segmentation algorithms as a proxy for image quality. Methods included a segmentation-based quality assessment framework for retinal images that used unsupervised vessel segmentation and then extracted features from the binary segmentation image for classifying the quality of an image using Support Vector Machine(SVM) and ensemble decision tree classifiers [97], pixel-wise binary segmentation based on AdaBoost with local texture descriptors to compared the Dice overlap coefficient with manual segmentations to a variety of quality metrics [98], and a

CNN to segment the macular region in eye fundus images for image quality assessment, where good image quality was when the macular area was above a threshold [99].

A limitation of segmentation-based quality assessment approaches is the lack of a direct relationship between segmentation accuracy and diagnostic image quality, where errors in segmentation may not be reflected in reduced diagnostic performance. The segmentation algorithm and evaluation metrics chosen can affect the result, which consequently makes the comparison of different methods difficult [100]. Many studies use relatively simple segmentation tasks or focus on particular anatomical structures that might not generalize well to more complicated clinical scenarios. Segmentation-based quality metric performance can also strongly depend on the imaging modality and anatomical region, which could limit their broad applicability [101]. While machine learning-based approaches for segmentation-based quality assessment show great promise, they require large annotated datasets to train, which can be challenging to obtain for medical images [102].

2.5 Pose estimation

Pose estimation has emerged as a computer vision task with far-reaching implications across various domains [103]. This technology aims to detect and localize anatomical keypoints or body parts of humans or objects within images or videos. The field has advanced in recent years, particularly with the advent of deep learning techniques that have revolutionized the accuracy and efficiency of pose estimation algorithms [104].

Pose estimation has two main categories: top-down and bottom-up. In top-down, posture inside each bounding box is independently estimated after person detection [105]. To achieve the expected outcomes, the subject of interest had to be located by the detection approach [106]. Bottom-up approaches connect each person to the body by first determining where joints or body segments are located [107].

Numerous pose estimation techniques such as Hyperpose [108], OpenPose [109], DeepLabCut [110], Posenet [111], DeepPose [112], Movenet Singlepose Lightning[113], Movenet Singlepose Thunder[114], Movenet Multipose Lightning[115] have been published in the past ten years. These methods make it possible to train new networks that are suited to particular research or clinical requirements or to use publicly available pre-trained networks. From the pose estimation keypoints, quality metrics can be calculated by post processing techniques for the desired

application. To compute the video quality score for gait videos, plane detection, zoom detection, multiple person detection, resolution of the person of interest are computed from the keypoints of pose estimation model. For real time applications, faster models provide almost instant quality score and feedback so that the video can be retaken by user.

2.5.1 Comparison of pose estimation models

This section compares three popular pose estimation models OpenPose, PoseNet, and MoveNet [116] (Table 2.2). MoveNet has three versions, Single Pose Lightning, Single Pose Thunder, and Multipose Lightning. PoseNet and MoveNet both identify 17 key-points, with 5 located on the face and 12 on the body. In contrast, OpenPose offers a more comprehensive tracking system, detecting a total of 137 keypoints: 25 on the body (including the feet), 21 on each hand, and 70 on the face. This extensive key-point detection allows OpenPose to capture body movements with greater precision and detail compared to the other two models. Another difference lies in the pose estimation method. OpenPose and MoveNet employ the bottom-up approach, whereas PoseNet employs the top-down approach.

Table 2.2 Specifications of OpenPose, PoseNet, MoveNet[116]

	OpenPose	PoseNet	MoveNet
Detection Parts	Body, Foot, Hand, Face	Body, Parts of Face	Body, Parts of Face
Keypoints	137	17	17
Method	Bottom-up	Top-down	Bottom-up

To compare the performance of OpenPose, PoseNet, MoveNet Lightning and MoveNet Thunder, Images from COCO [117] and MPII datasets [118] were grouped into two groups. Group 1 contained person and Group 2 was without any person. Each group had 1000 images.

Table 2.3 summarises the accuracies on images of two groups. PoseNet had the highest accuracy and Movenet Lightning had the lowest accuracy. The average accuracy across single-person and no-person images was 97.6% for PoseNet, 86.2% for OpenPose, 80.6% for MoveNet Thunder, and 75.1% for MoveNet Lightning. This suggests that PoseNet might be the preferred choice for applications prioritizing accuracy over speed.

Table 2.3 Accuracy of OpenPose, PoseNet, MoveNet Lightning, and MoveNet Thunder [116].

Model	OpenPose	PoseNet	MoveNet Lightning	MoveNet Thunder
With person	78.5%	96.7%	78.7%	79.5%
Without person	93.85%	98.4%	71.5%	81.6%
Average	86.25	97.6%	75.1%	80.6%

In terms of processing speed, MoveNet Lightning had the fastest performance, followed by PoseNet, MoveNet Thunder, and OpenPose. The average time to estimate poses in 1,000 images was 53.458 seconds for MoveNet Lightning, 75.232 seconds for PoseNet, 139.974 seconds for MoveNet Thunder, and 643.536 seconds for OpenPose.

The comparison revealed that MoveNet Lightning was the fastest model. OpenPose was 12 times slower than MoveNet and 8.5 times slower than PoseNet, but only OpenPose could estimate the poses of multiple persons. In 2023, Google released a Movenet Multipose Lightning version that can detect up to 6 persons. Google claimed Movenet Multipose Lightning is 1.6 times slower than the Movenet Lightning version [113], [115] and the accuracy is same as Movenet Lightning. With this, Movenet Multipose Lightning should detect 1000 images in 83.16 seconds, which is 7.5 times faster than the OpenPose.

2.5.2 MoveNet Multipose Lightning architecture

A Google-based inference model called MoveNet [119] was created by the digital health startup IncludeHealth. The convolutional neural network model runs on RGB images and predicts human joint locations of people in the image frame. The main differentiator between this MoveNet MultiPose and its precedent, MoveNet SinglePose model, is detecting multiple people in the image frame at the same time while still achieving real-time speed [113], [114], [115]. Movenet Multipose Lightning uses MobileNetV2 image feature extractor with Feature Pyramid Network [120] decoder followed by CenterNet [121] prediction heads with custom post-processing logics. Lightning uses depth multiplier 1.5 [115].

MoveNet is a sophisticated bottom-up pose estimation model that employs heatmaps to accurately localize human keypoints. Its architecture is composed of two primary components: a feature extractor and a set of prediction heads. The model's design draws inspiration from CenterNet [121], but incorporates modifications to enhance both speed and accuracy.

MobileNetV2 [122] is at the core of MoveNet's feature extraction process, augmented with a feature pyramid network (FPN) [120]. This combination enables the model to generate high-resolution feature maps with rich semantic information, utilizing an output stride of 4. The feature extractor is complemented by four distinct prediction heads, each serving a crucial role in the pose estimation process.

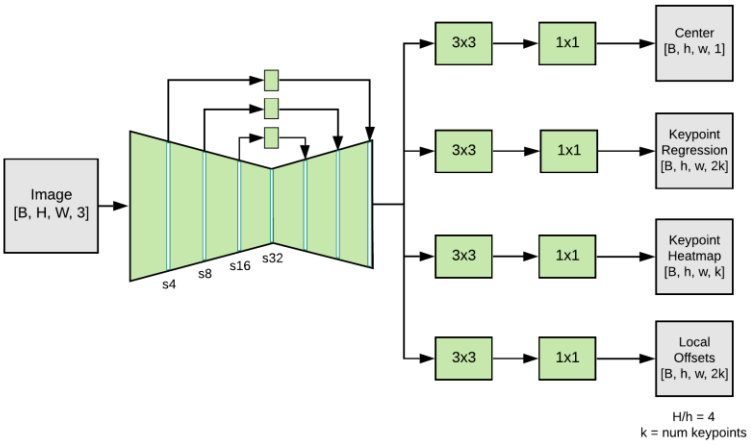

Figure 2.2 MoveNet Architecture [119]

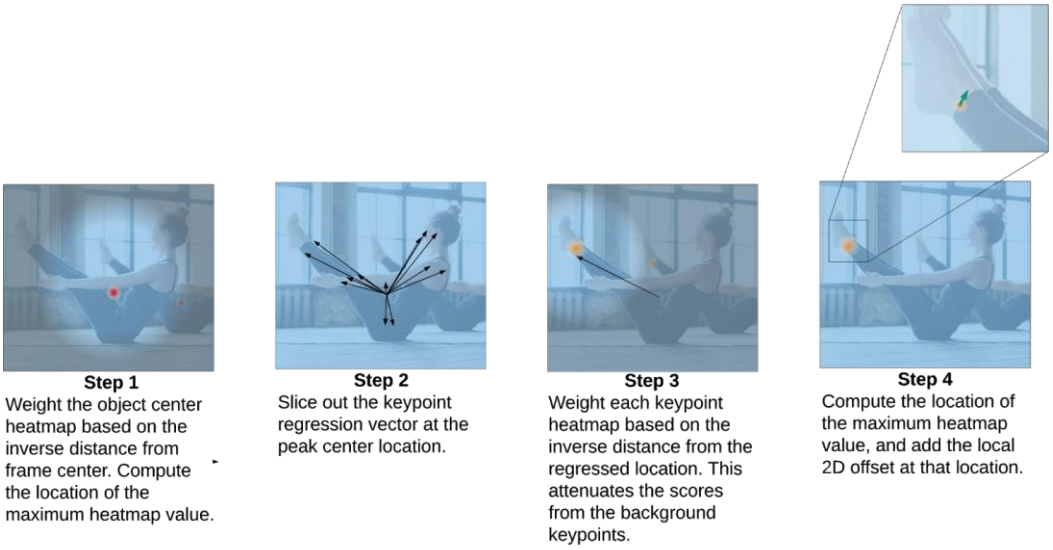

Figure 2.3 MoveNet post-processing steps [119]

The first prediction head generates a person center heatmap, which identifies the geometric centers of individuals within the frame. The second head produces a keypoint regression field, predicting a complete set of keypoints for each person and facilitating the grouping of keypoints into distinct instances. The third head creates a person keypoint heatmap, localizing all keypoints independently of person instances. Finally, the fourth head generates a 2D person-keypoint offset field, refining the precise sub-pixel location of each keypoint relative to the output feature map pixels.

This architectural design allows MoveNet to efficiently process input images and produce accurate pose estimations, making it a powerful tool for various applications in computer vision and human-computer interaction [119].

2.6 Conclusion

The literature review reveals many advancements in gait analysis, automated EVGS, and video quality assessment techniques. The EVGS has emerged as a valuable tool for assessing gait abnormalities, with recent efforts focusing on automating the scoring process to improve efficiency and reduce subjectivity. However, the success of automated EVGS systems heavily relies on the quality of input videos, highlighting the need for robust video quality assessment methods.

Content-based video quality assessment has evolved to incorporate both technical and semantic elements, recognizing that human perception of quality extends beyond mere technical parameters. In the context of medical image/video applications, task-based quality assessment methods have shown promise in evaluating the effectiveness of images and videos for specific clinical tasks [4].

Pose estimation techniques have progressed, with models like OpenPose, PoseNet, and MoveNet offering various trade-offs between accuracy and speed. Among these, MoveNet Multipose Lightning stands out as a promising solution for real-time multiple person detection, offering a balance between speed and accuracy [116].

Given the requirements for automated EVGS and the need for efficient, accurate pose estimation, MoveNet Multipose Lightning appears to be an excellent choice for extracting keypoints and developing quality metrics for video analysis. Its ability to detect multiple persons quickly while maintaining reasonable accuracy aligns well with the goals of automated gait analysis [115]. By leveraging MoveNet Multipose Lightning for pose estimation and developing

custom objective video quality assessment metrics, it becomes possible to create a system that can efficiently analyze patient videos and classify them as suitable or unsuitable for input into automated EVGS algorithms. This approach combines the strengths of advanced pose estimation techniques with tailored quality assessment methods, potentially leading to more reliable and efficient gait analysis in clinical settings.

3 Methodology

3.1 Overview

This section describes a structured methodology for automatic video quality assessment of gait video that includes pose estimation techniques, evaluation metrics, and automatic quality scoring mechanisms. In this approach, improper plane, low-resolution, multiple and overlapping people, abnormal zooming are identified.

The algorithm was developed and validated using a dataset of videos from the Sanatorio del Norte medical center in Tucumán, Argentina. This dataset included gait videos in sagittal, coronal, and transverse planes of 230 individuals with walking disorders. Videos were recorded at 60 Hz and captured in a closed environment with good lighting. The dataset was divided into development and testing sets, with 161 videos in the development set and 69 videos in the testing set. The first 161 videos in the overall dataset were allocated to development

The overall video quality analysis process involves:

1. Loading video in the system and using MoveNet Multipose Lightning to detect all the persons and extract their skeletal keypoints.
2. Check for multiple person detection in the video.
    a. If more than one person is detected, user input is required to identify the desired person to ensure that the appropriate individual is processed in succeeding steps.
3. The extracted keypoints are used to calculate metrics for video quality assessment.
4. Check for overlapping individuals, flagging such videos for manual review or re-recording.
5. Plane detection classifies the video into sagittal, coronal, transverse views to ensure proper orientation.
6. Zoom detection evaluates sudden changes in the person's bounding box height, which may signify unintentional zooming.
7. Resolution evaluation ensures that bounding box dimensions are sufficient for pose estimation algorithm to clearly detect keypoints.
8. Quality scoring assigns an overall quality score to the video. From the evaluation metrics, a Random Forest classifier classifies whether the video is good or requires manual modification.
    a. Videos that fail the quality check can trigger a feedback mechanism with recommendations for re-recording.

3.2 Multiple people

Output from MoveNet Multipose Lightning provides bounding boxes for each person in the video. Therefore, we know the number of people in the video by counting the number of bounding boxes. A bounding box is a rectangular area that covers the person, defined with x1, y1, x2, y2 coordinates. The pose estimation model also assigns a confidence score to each detected person in the frame, which indicates the detection quality. Detections with a confidence threshold less than the defined minimum value (default value is 0.5) are considered lower-reliability detections. The lower reliability detections are eliminated from further analyses.

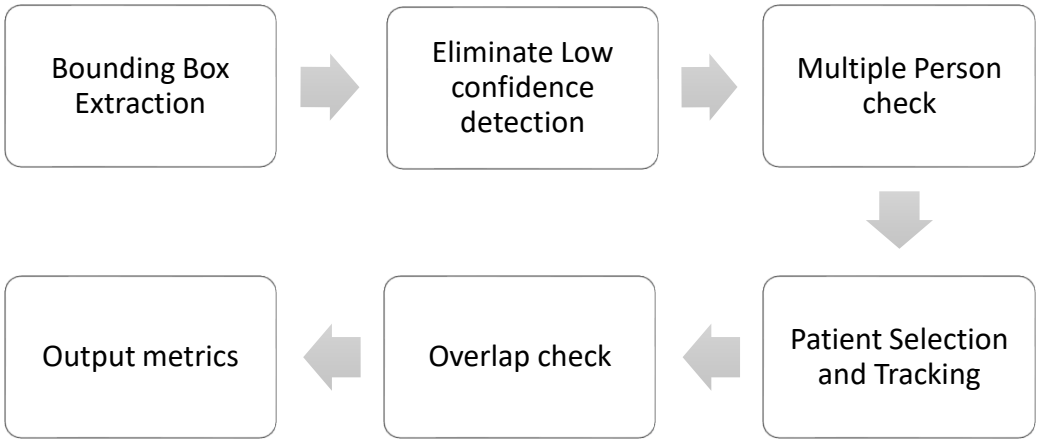

Figure 3.1 Steps to calculate multiple persons and overlap score

If the algorithm finds more than one person, it prompts the user to select the person of interest in that frame. This selection is made possible through a graphical user interface (GUI) where the user can see the frame with overlaid bounding boxes and keypoints for each detected person (Figure 3.2). If the frame does not show the target person, the user has the option to skip to the next frame with more than one detected person.

(a) (b)

Figure 3.2 Graphical user interface to (a) select patient, (b) track the selected patient

Once selected, the chosen person's keypoints are checked in future frames to ensure that the same person is selected from the current frame to end of video. A similarity measure based on the Euclidean distance between the centroid of the target person's bounding box in the previous frame and the centroid of all bounding boxes in the current frame is used to determine the closest match.

$$d = \min\left(\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2}\right), \forall i \in \{1,2,\dots,n\} \quad (3.1)$$

In equation 3.1, (x_t,y_t) is the centroid of target patient in previous frame , (x_i,y_i) is the centroid of the i^{th} bounding box of current frame, n represents number of people detected by pose estimation model and d represents the Euclidean distance of centroid between target patient's bounding box in previous and current frames. The bounding box in the current frame whose centroid has the smallest Euclidean distance from the centroid of the target patient's bounding box in the previous frame is considered to be the target patient's bounding box in the current frame. As a visual aid, the selected person has a green bounding box to differentiate from other detected people in the frame (Figure 3.2).

The multiple person quality metric is the percentage of the number of frames with only the target patient's bounding box, relative to the total number of frames.

Overlap are frames with the target person is occluded by another person in the video. The conditions to check overlap are:

- x_1 of the target person bounding box is less than x_2 of any other bounding box in the frame
- x_2 of the target person bounding box is greater than x_1 of any other bounding box in the frame
- y_1 of the target person bounding box is less than y_2 of any other bounding box in the frame
- y_2 of the target person bounding box is greater than y_1 of any other bounding box in the frame

Note: x_1 is the left edge, y_1 is the top edge, x_2 is the right edge, and y_2 is the bottom edge of the bounding box.

If one of the above conditions is true, then the frame has overlapping bounding boxes and this frame is flagged as an overlapping frame. Overlap detection is repeated for all the frames in the video. The quality metric for multiple people overlap is the ratio (percentage) of number of overlap frames to the total number of frames.

3.3 Plane detection

Plane identification is a very important aspect of gait video quality assessment, since outcome measures are typically related to either sagittal or coronal planes. Automated EVGS calculates 12 EVGS parameters from the coronal plane and 5 EVGS parameters from the sagittal plane [3]. The plane detection algorithm classifies each video as either a sagittal, coronal, or transverse plane based on the geometric orientation of the human body (Figure 3.3).

The first step in plane detection is to extract left shoulder, right shoulder, left hip, and right hip keypoints in every frame of the video. To assess the video perspective and determine the plane of observation, three geometric ratios are derived based on the selected keypoints.

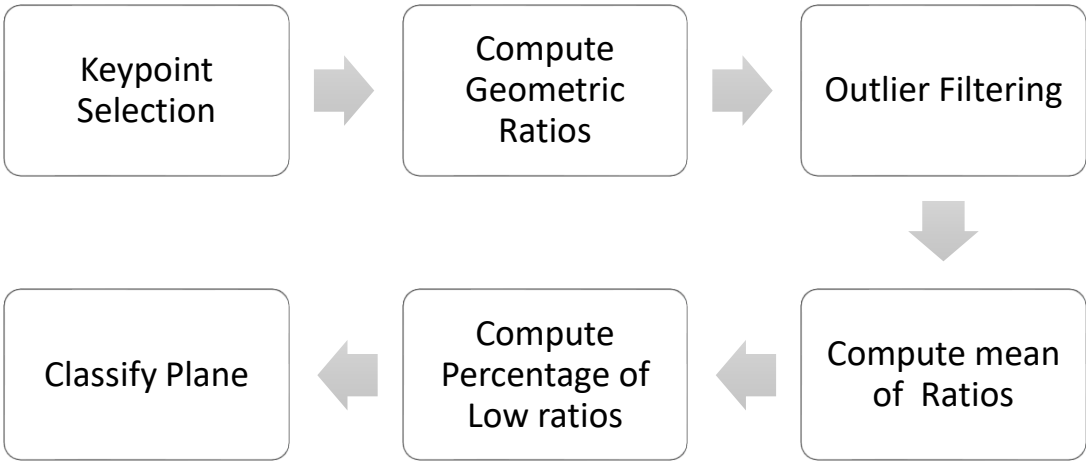

Figure 3.3 Steps to classify the video plane

Horizontal Shoulder Distance (HSD): The horizontal shoulder distance is a measure of the width of the upper body in the frame, measured across the horizontal distance between the left and right shoulders. This gives insight into the orientation or alignment of the body in the horizontal plane and is computed as

$$HSD = \left| x_{left_shoulder} - x_{right_shoulder} \right| \tag{3.1}$$

Right Vertical Shoulder-to-Hip Distance (RVSD): Right Vertical Shoulder-to-Hip Distance refers to the vertical positioning of the right side of the body. This involves the vertical area between the right shoulder and the right hip and represents the posture of the right side in relation to the frame. This is computed as

$$RVSD = \left| y_{right_shoulder} - y_{right_hip} \right| \tag{3.2}$$

Left Vertical Shoulder-to-Hip Distance (LVSD): The Left Vertical Shoulder to Hip Distance represents the vertical positioning of the left side of the body. This distance is between the left shoulder and left hip vertically, reflecting the posture of the left side within the frame. This is computed as

$$LVSD = \left| y_{left_shoulder} - y_{left_hip} \right| \tag{3.3}$$

Additional metric are also computed. The Initial Ratio (r1) is the ratio between HSD and RVSD, written as

$$r1 = \frac{HSD}{RVSD} \tag{3.4}$$

The Second Ratio (r2) ishe ratio between LVSD and HSD, given as

$$r2 = \frac{LVSD}{HSD} \tag{3.5}$$

The Final Ratio (R) is the ratio of r2 and r1 found from

$$R = \frac{r2}{r1} \tag{3.6}$$

The geometrical ratios are essential to understand the body orientation and to differentiate between sagittal and coronal views.

Outlier Filtering

Videos where the person is walking in one direction can be classified using *R* (i.e., right-to-left or left to right in sagittal videos and towards camera or away from camera in coronal videos). However, if the is person walking in the sagittal plane, *R* changes when they turn around. This might lead to misclassification. In addition, the ratios computed over all frames may have extreme values contributed by either noise or poorly detected keypoints. Therefore, to make the method robust, outlier filtering was implemented by applying the interquartile range.

- The first step is calculation of 25th (P25) and 75th (P75) percentiles of the ratio list
- Then, ratios outside the P25 to P75 range are excluded

The ratio list without outliers is termed the Filtered ratio list and the mean of the Filtered ratio list is used to classify the general body orientation of the video.

Plane Identification

Values in the Filtered ratio list that are less than 0.7 are termed as “low ratios”. The threshold of 0.7 was determined empirically through a grid search optimization process, ensuring it effectively distinguishes sagittal, coronal, and transverse views across diverse video datasets. The percentage of low ratios in the filtered ratio list was calculated by summing the number of frames with R below 0.7 and dividing this by the total number of frames in the list.

The plane of the video can be classified as sagittal, coronal or transverse, as follows:

Sagittal plane: The mean of filtered ratio list is below 0.5, or mean of filtered ratio list is between 0.5 to 1 and the percentage of low ratios is more than 40. This condition suggests the body was oriented sideways relative to the camera position.

Coronal Plane: The mean of filtered ratio list is between 0.5 and 1 and the percentage of low ratios is less than 40 or the mean of filtered ratio list is between 1 to 2 or the mean of filtered ratio list is between 2 to 3 and the percentage of low ratios is not equal to 0. This condition suggests orientation of the body is in front or backside facing the camera.

Transverse Plane: The mean of filtered ratio list is 2 or more, or the mean of filtered ratio list is between 2 to 3 and the percentage of low ratios is equal to 0. This condition indicates the plane is transvere.

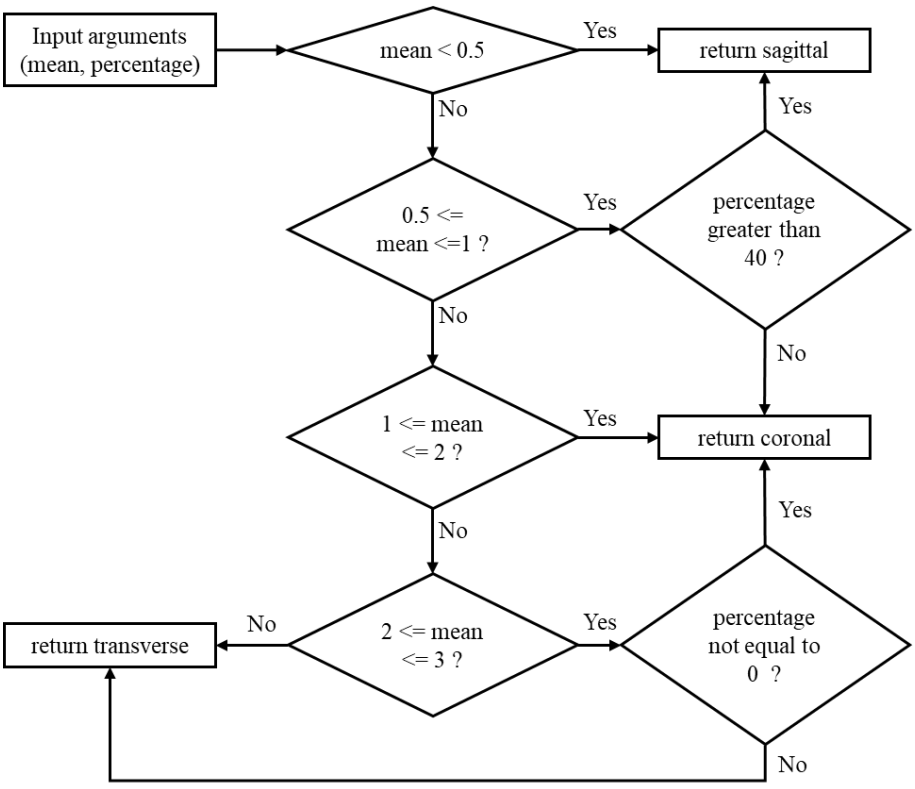

Figure 3.4 Classification criteria for plane classification

Figure 3.4 shows the classification flow chart for plane identification. Mean refers to the average of the filtered ratio list. Percentage indicates the percentage of low ratios in filtered ratio list.

3.4 Zoom detection

Zoom detection is important since abrupt changes in the scale may introduce pose estimation and parameter errors. This section identifies sudden changes in bounding box height around the person of interest.

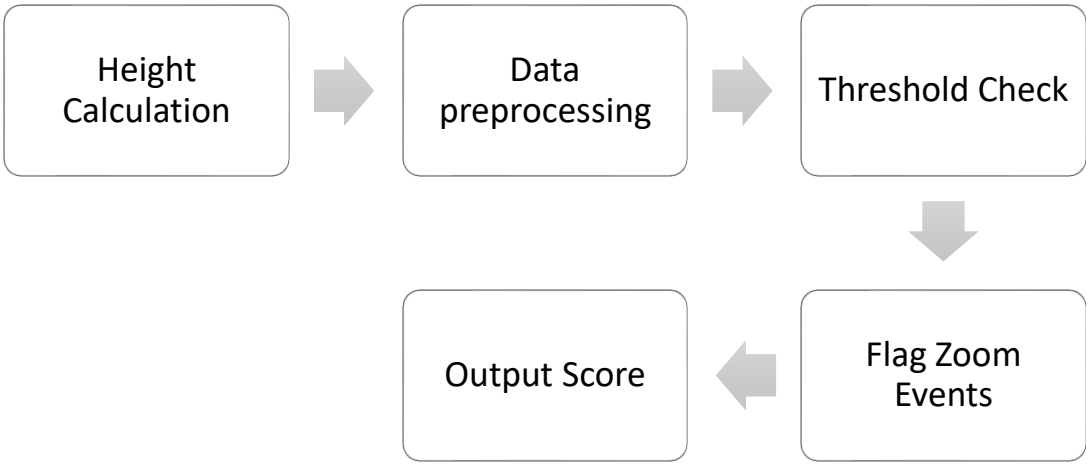

Figure 3.5 Zoom detection flowchart

The bounding box coordinates (x1, y1, x2, y2) are used to calculate bounding box height:

$$height = y2 - y1 \tag{3.7}$$

Since video resolutions can be very different, raw height is normalised by video frame height. This ensures the algorithm will be resolution-independent, thus capable of handling videos of any aspect ratio and resolution.

$$normalised\ height = \frac{height}{frame\ height} \tag{3.8}$$

Height data is usually noisy because of minor inaccuracies in pose detection or slight postural changes by the person. A Gaussian filter is thus applied to the normalized height data to reduce noise. Data smoothing depresses the noise without affecting major trends, providing a clean height profile suitable for sharp variation detection and to prevent false positives.

$$smoothed_height = gaussian_filter\left(normalised_{height}, \sigma\right) \quad (3.9)$$

where σ is the standard deviation of the Gaussian kernel. Here, σ = 2 is adopted to achieve a trade-off between noise reduction and trend preservation.

Once the algorithm has smoothed height data, height changes between frames are calculated based on fixed step sizes. The smoothed height difference for every i^{th} frame is given by:

$$\Delta height = smoothed\ height_{i+s} - smoothed\ height_i \quad (3.10)$$

The step-size difference approach ensures that only relevant changes in height are considered and, in this way, highlights major zoom events rather than gradual fluctuations. In the implementation, where video frame rate is 60 Hz, s=8 provides a good trade-off between sensitivity and robustness.

The height differences are then compared to a threshold value τ = 0.05 to determine if a frame represents a zoom event. A Zoom-In event is identified when $\Delta height > \tau$, since height increases suddenly. Frames with $\Delta height < -\tau$ are defined as Zoom-Out due to the fast decrease in height.

3.5 Video segmentation

This section explains how the algorithm systematically segments and scores usable video segments for gait analysis, considering both coronal and sagittal plane. Extracted video frames ensure that no zoom event has occurred and that the video segment contains at least 2 walking strides. Each plane is analyzed independently to account for the unique characteristics of the respective planes.

3.5.1 Extract video segments without zoom from coronal plane video

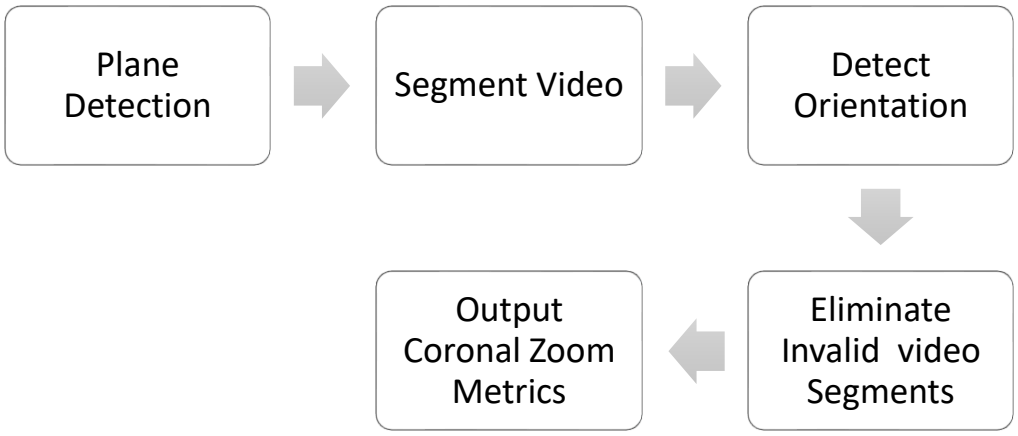

Figure 3.6 Steps to extract no zoom video segments in coronal plane

The algorithm starts with segmenting the input video based on the change in the person's height (Figure 3.6). Segmentation is based on detecting sign changes in the height difference as given in

$$\Delta h_t = h_t - h_{t-1} \tag{3.11}$$

Here, h_t is the height of the person in current frame t, h_{t-1} is the height of the person in previous frame t-1, Δh_t represents change in height.

A new segment is created when sign of Δh_t is not equal to sign of Δh_{t-1}. That is, one segment would be generated with every transition from an increasing to a decreasing height or decreasing to increasing height. For the videos without zoom, a segment is generated only when the person's position relative to the camera changes. Specifically:

- If the person is walking away from the camera, a new segment is created when they turn and start walking toward the camera.
- Similarly, if the person is walking toward the camera, a new segment is created when they turn and start walking away from the camera.

However, if there is a zoom event, then a segment will also be generated when zooming in (patient height will increase) and zooming out (patient height will decrease). For each segment, the algorithm detects the person's orientation to the camera. Orientation here refers to the direction the person is facing, whether they are facing toward the camera (front) or away from the camera (back).

The orientation is decided by shoulder alignment analysis, using the x-coordinates of the right and left shoulder keypoints. When the left shoulder keypoint is to the right of the right shoulder keypoint (i.e., left shoulder > right shoulder x-coordinate), the person is oriented with their front towards the camera. For orientation back to the camera, the left shoulder keypoint would be to the left of the right shoulder keypoint (i.e., left shoulder < right shoulder x-coordinate). In general, when the person is walking towards the camera, their height will gradually increase and when the person walks away from camera their height will decrease. The algorithm eliminates video segments with a zoom event, and the video segments with zoom can be identified by the following conditions:

- Person walking towards the camera and height of the person is decreasing.
- Person walking away from the camera and height of the person is increasing.

In addition to eliminating video segments with zoom, video segments with duration less than three seconds, an approximation for the time used to complete two strides, were rejected since two strides is the minimum requirement for automated EVGS analysis. After eliminating video segments with less than three seconds and video segments with zoom, the remaining video segments are considered as valid video segments. If the number of frames between two valid video segments is less than 5 frames (i.e., difference of end frame number of a valid video segment and start frame number of the next video segment is less than 5), the two video segments are merged into one valid video segment. The merged valid video segment has a start frame number from the first video segment and end frame number from the next valid video segment. This merging method maintains the continuity of the gait cycle. These methods ensure that only video segments suitable for automated gait analysis are preserved.

3.5.2 Calculate score for video segments

All valid video segments are scored for quality, with scores ranging from 0 to 100. The duration of each valid video was extracted. Moderate intensity walking is 100 steps/minute; therefore, the approximate time to complete five steps is 8 seconds (5 steps ensures that 2 viable strides are available for analysis). So, video segments with a minimum 8 second duration are awarded a maximum quality score (100) and video segment of 3 seconds or less are scored as 0. For video durations between 3 and 8 seconds, the score is scaled proportionally from 0 to 100 based on the segment's length relative to the 0–8 second scale.

Two of the highest scores are selected: one related to the segments where the person is oriented towards the camera, and another where the person is oriented away from the camera. Finally, general video quality is scored by computing the percentage of frames that belong to valid segments, with respect to the total number of frames in the video.

$$Percentage\ of\ valid\ frames = \frac{Total\ no\ of\ frames\ in\ valid\ video\ clips}{Total\ no\ of\ frames\ in\ the\ video} \tag{3.12}$$

Two highest score of each orientation and percentage of valid frames help to assess extracted video segments for further automated gait video analysis in the coronal plane.

3.5.3 Extract video segments without zoom from sagittal plane video

To classify video segments in the sagittal orientation (i.e., person walking from left-to-right or right-to-left), x-coordinates of the ears and nose keypoints are used. Since the coordinate system for the image starts from the top-left, with increasing values to the right for the x-coordinate:

- If the x-coordinate of the ears is greater than the nose x-coordinate, the person is walking from right-to-left (i.e., ears are always to the right of the nose).
- If the x-coordinate of the ears is less than the nose x-coordinate, the person is walking left-to-right walking (i.e., ears are left of the nose).

One video segment is generated for every transition from left-to-right to right-to-left, or from right-to-left to left-to-right. In each video segment, the person is either walking left-to-right or right-to-left, not both.

Video segments are analyzed to detect zoom events. Unlike the coronal plane, the height of the person in the sagittal plane remains relatively constant, since the individual stays a similar distance from the camera. A zoom event in a video segment is identified using same approach discussed in chapter 3.4. As described in equation 3.10, Δheight represents the difference in the height of the person's bounding box between frame i and frame i+8. If Δheight exceeds 0.05, the frames from i to i+8 are identified as zoom event frames and are eliminated. This elimination process is repeated at all instances where Δheight > 0.05, removing all zoom events and resulting in video segments free of zoom events.

After this process, video segments with duration less than three seconds, an approximation for the time used to complete two strides, were rejected since two strides is the minimum

requirement for automated EVGS analysis. After removing video segments with duration less than 3 seconds, the remaining video segments are considered as valid video segments

3.5.4 Calculate score for sagittal video segments

Scoring for the sagittal plane videos follows the same method as for the coronal plane videos. Each video segment is scored based on duration of the video (section 3.5.2 contains the detailed methodology). Scoring to assess zoom in sagittal video includes highest score of left-to-right video segments, highest score of right-to-left video segments, and percentage of valid frames in the video.

3.6 Video quality classification and feedback

3.6.1 Overall quality classification

The objective of the quality classification is to classify the gait video as “good”, “manual edit required”, or “in the transverse plane”. Good videos can directly be used by automated EVGS algorithms and manual edit required videos need to edit (trimming and/or cropping) before they can be used in the automated EVGS algorithm.

Since automated EVGS processes only sagittal and coronal planes, videos taken from transverse plane are identified using plane detection methodology discussed in section 3.3. If the video is identified as transverse plane, then the video is excluded from automated EVGS processing.

A Random Forest classifier is used to classify the coronal and sagittal videos as “good” or “manual edit required” videos. Gait videos from the development dataset were used to train the Random Forest classifier and videos from the validation dataset were used to validate Random Forest classifier performance. Quality scores are calculated as explained in sections 3.2 to 3.5 for all the videos in the development dataset. After this, the classification framework uses the calculated scores as input parameters for the random forest classifier: multiple persons (percent of frames with multiple people), multiple persons overlap (percent of frames with overlapped bounding boxes), zoom detection (1 if zoom occurred, 0 if no zoom), extracted clip without zoom (maximum score for strides in a video segment), and resolution (percent of frames with acceptable resolution).

Hyperparameters such as number of estimators, maximum depth, minimum samples required to split a node, minimum number of samples required at leaf node and bootstrap sampling of

random forest classifier were tuned using the grid search approach from scikit-learn. The grid search approach calculates performance metrics such as accuracy, precision, sensitivity, specificity, F1 score for all combinations of hyperparameters and returns the best hyperparameters. The selected hyperparameters for the random forest classifier are listed in the Table 3.1

Table 3.1 Fine-tuned hyperparameters using grid search approach

Hyperparameters	**Fine tuned value**
number of estimators	50
maximum depth	None
minimum samples required to split a node	2
minimum samples required at leaf node	1
bootstrap	False

A Random Forest classifier classifies the input parameters by selecting the class with highest probability, using the input parameters from the development dataset videos random forest classifier trained with the best hyperparameters. The Random Forest classifies the class with highest probability. The probability value of class “Good videos” is used as the overall quality score. Probability is in the scale between 0 and 1. The overall quality score is defined as 0 to 100, for uniformity with the previously calculated metrics scores and user-friendly interpretation. This scale conversion is achieved via

$$Overall\ quality\ score = 100 * Probability\ of\ "Good\ videos" \quad (3.13)$$

Video overall quality scores greater than 50 are considered good videos. All other videos are considered as manual editing required.

3.6.2 Quality feedback generation

Providing quality feedback to users is a vital step in improving video quality for gait analysis. This methodology ensures that users receive actionable suggestions, enabling continuous improvement of video quality. Feedback is based on individual metric scores and helps users identify and address specific issues in their recordings. If any metric score falls below 80, the algorithm provides targeted suggestions for improvement.

- **Plane Detection**: if a sagittal or coronal view is not identified
    - Suggestion for sagittal plane: "Ensure the camera is positioned above the pelvis height and parallel to the walkway (at a 90-degree angle to the side of the body)."
    - Suggestion for coronal plane: "Ensure the camera is at hip height and in front of the person."
- **Multiple person score**: if multiple persons are detected in the video, feedback recommends recording with only the person of interest in the frame.
    - Suggestion: "Ensure only the person of interest is in the frame during recording. "
- **Multiple persons overlap score**: if overlapping individuals are detected, then the video has more than one person is in the frame. Feedback advises avoiding multiple persons in the frame.
    - Suggestion: "Ensure only the patient person of interest is in the frame during recording. Remove additional persons and avoid overlap if more than one person is required."
- **Zoom detection**: if zoom artifacts are detected, feedback prompts users to maintain a stable camera.
    - Suggestion: "Keep the camera steady without zooming and stay the same distance from the person."
- **Videoclip length without zoom score**: video is too short
    - Suggestion: "Ensure the person walks for at least 2 strides (4 steps). Ideally the person was walk 5-6 strides, with the middle strides captured on video."
- **Resolution score**: if the resolution score of detected person is lower than 60*80 pixel, feedback suggests to adjust the camera distance to increase the number of pixels that define the person in the frame.
    - Suggestion: "Ensure the person occupies at least 50 percentage of total frame height. Adjust the camera distance accordingly."

4 Validation

Validation of the quality algorithms is a critical step in determining effectiveness of the algorithm performance with unseen real-world data. The dataset was divided into development and validation datasets. The development dataset is used for developing and optimizing the algorithms. Once the algorithm gives expected output consistently on the development dataset, the validation dataset is used to evaluate the algorithm on unseen data.

Algorithm performance was evaluated using accuracy, precision, sensitivity, specificity, and F1-score, and Matthews Correlation Coefficient (MCC). Accuracy is the proportion of all classifications that were correct, whether positive or negative [123],

$$Accuracy = \frac{corrrect\ classifications}{total\ classifications} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

The percentage of all the model's positive classifications that are truly positive is known as precision [123].

$$Precision = \frac{correctly\ classified\ actual\ positives}{everything\ classified\ as\ positive} = \frac{TP}{TP + FP} \tag{4.2}$$

The percentage of all true positives that were correctly classified as positives is known as sensitivity [123].

$$Sensitivity = \frac{correctly\ classified\ actual\ positives}{all\ actual\ positives} = \frac{TP}{TP + FN} \tag{4.3}$$

The percentage of all true negatives that were correctly classified as negatives is known as specificity [123].

$$Specificity = \frac{correctly\ classified\ actual\ negatives}{all\ actual\ negatives} = \frac{TN}{TN + FP} \tag{4.4}$$

F1 score is the harmonic mean of precision and recall [124].

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.5}$$

The MCC scale, which goes from -1 to +1, represents the quality of binary (two-class) classifications. A perfect forecast has a score of +1, a random prediction has a score of 0, and an inverse prediction has a value of -1 [125].

$$MCC = \frac{TP + TN - FP * FN}{\sqrt{(TP + FP)(TP + FN\ \ )(TN + FP)(TN + FN)}} \tag{4.6}$$

4.1 Multiple person detection

The validation dataset for evaluating the Multiple Person Detection algorithm had 58 videos that contained 38 videos with only one identifiable person for the entire video, and 20 videos that contained two or more persons at the same time. A graduate student manually reviewed each video to classify as single person or multiple persons. Algorithm performance was assessed by comparing algorithm results with the manual classification. A confusion matrix was formed that included:

- True Positives (TP): Videos correctly identified as containing multiple people.
- False Positives (FP): Videos falsely identified as containing more than one person.
- True Negatives (TN): Videos correctly identified as containing a single person.
- False Negatives (FN): Videos where the presence of multiple persons was missed.

4.2 Multiple persons overlap detection

The validation dataset for this validation was categorized into two sets: a set of 15 videos with overlap of one person over another person (overlap present set), and a set of 43 videos without overlap of one person over another person (overlap absent set). Each video was manually reviewed and classified as overlap present or overlap absent. Algorithm performance was assessed by comparing algorithm results with the manual classification. A confusion matrix was formed that included:

- True Positives (TP): Videos correctly identified as containing overlaps.
- False Positives (FP): Videos incorrectly flagged as having overlaps.
- True Negatives (TN): Videos correctly identified as containing no overlaps.
- False Negatives (FN): Videos where overlaps were not detected despite their presence.

4.3 Plane detection

The validation dataset for plane detection had 46 videos in which 19 videos were coronal, 15 were sagittal, and 12 were transverse. Each video was manually reviewed to classify as coronal or sagittal or transverse plane and this manual classification was used as ground truth to validate algorithm classification performance.

4.4 Zoom detection

The validation dataset for evaluating the zoom detection had 26 videos in which 14 videos had a zoom event (i.e., zooming-in or zooming-out) and 12 videos without zoom events. Each video in the dataset was manually classified as having a zoom event or without a zoom event. Manual classification was used as ground truth to validate the identification of zoom events algorithm. A confusion matrix was formed that included:

- True Positives (TP): Videos correctly identified as containing zoom event.
- False Positives (FP): Videos incorrectly flagged as having zoom event.
- True Negatives (TN): Videos correctly identified as containing no zoom event.
- False Negatives (FN): Videos where zoom event was not detected despite their presence.

4.5 Extract video segment without zoom event

The validation dataset for evaluating the algorithm that extracts video segments without zoom event consists of videos with zoom event and videos without zoom event. First the algorithm extracted two video segments. For videos from coronal plane, one video segment for person walking towards the camera and one video segment for person walking away from camera. For sagittal plane videos, one video segment for person walking left-to-right and one video segment for person walking right-to-left. The extracted video segments were manually checked for presence of zoom event, video segments of each walking orientation for every video, and minimum duration of video segment is greater than 3 seconds. If there was a zoom event, then the video segment was flagged. This algorithm was evaluated using accuracy (equation 4.7).

$$Accuracy = 1 - \left(\frac{No\ of\ flagged\ video\ clips}{Total\ No\ of\ extracted\ video\ clips}\right) \quad (4.7)$$

4.6 Overall video quality classification

The final Video Quality Assessment framework was evaluated using a validation dataset that had 20 videos classified as “Good videos” and 24 videos as “Manual edit required”. Good Videos can be used in automated EVGS algorithm without any changes and manual edit required videos cannot be used in automated EVGS algorithm without editing the video. Each video in the dataset was manually reviewed to classify video as good videos or manual edit required videos. Random Forest classifier performance was assessed by comparing algorithm results with the manual classification. A confusion matrix was formed that included:

- True Positives (TP): Videos correctly identified as “Good videos”.
- False Positives (FP): Videos incorrectly identified as “Good videos”.
- True Negatives (TN): Videos correctly identified as “Manual edit required videos”.
- False Negatives (FN): Videos incorrectly identified as “Manual edit required videos”.

5 Results

5.1 Multiple person detection

The results on the validation dataset show 3 misclassifications (Table 5.1, Table 5.2), but all outcome measures were greater than 0.90.

Table 5.1 Confusion matrix of multiple person detection for validation dataset

	True Multiple persons Detected	**True** Multiple persons **Not** detected
Predicted Multiple Person Detected	18	1
Predicted Multiple Person **Not** Detected	2	37

Table 5.2 Performance metrics for multiple person detection for validation dataset

Accuracy	Precision	Sensitivity (Recall)	F1 Score	Specificity	MCC
0.95	0.95	0.90	0.92	0.97	0.88

5.2 Multiple persons overlap detection

The performance of the algorithm for detecting overlapping persons on the development dataset is summarized Table 5.3 and Table 5.4. The algorithm had strong performance, with high recall (0.94) and specificity (0.95). Precision was the only score below 0.90. Only 3 videos were misclassified.

Table 5.3 Confusion matrix of multiple persons overlap detection for validation dataset

	True Multiple persons overlap Detected	**True** Multiple persons overlap **Not** detected
Predicted Multiple Person overlap Detected	14	2
Predicted Multiple Person overlap **Not** Detected	1	41

Table 5.4 Performance metrics for multiple persons overlap detection for validation dataset

Accuracy	Precision	Sensitivity (Recall)	F1 Score	Specificity	MCC
0.95	0.88	0.94	0.91	0.95	0.87

5.3 Plane detection

As shown in Table 5.5 and Table 5.6, plane detection worked for all validation set classifications.

Table 5.5 Confusion matrix of plane detection for validation dataset

	Actual Sagittal	Actual Coronal	Actual Transverse
Predicted Sagittal	15	0	0
Predicted Coronal	0	19	0
Predicted Transverse	0	0	12

Table 5.6 Performance metrics for plane detection for validation dataset

Accuracy	Precision	Sensitivity (Recall)	F1 Score	Specificity	MCC
1.0	1.0	1.0	1.0	1.0	1.0

5.4 Zoom detection

Zoom detection algorithm results are provided in Table 5.7 and Table 5.8. The algorithm had consistent results that ranged between 0.92 and 0.93. Only 2 videos were misclassified.

Table 5.7 Confusion matrix of zoom detection for validation dataset

	True Zoom Detected	**True** Zoom **Not** detected
Predicted Zoom Detected	13	1
Predicted Zoom **Not** Detected	1	11

Table 5.8 Performance metrics for zoom detection for validation dataset

Accuracy	Precision	Sensitivity (Recall)	F1 Score	Specificity	MCC
0.92	0.93	0.93	0.93	0.92	0.85

5.5 Extract video segments without zoom

The accuracy of the algorithm is summarized in Table 5.9. The algorithm demonstrated good performance on the validation dataset, processed 30 videos and extracted 60 video segments. Of these, only 5 clips were flagged, resulted in an accuracy of 92% in identifying video segments without zoom.

Table 5.9 Accuracy of the algorithm in the validation dataset

No of videos	No of video segments extracted	No of flagged video segments	Accuracy
30	60	5	0.92

5.6 Overall video quality classification

The Final Video Quality Assessment framework was evaluated to assess its ability to classify gait videos as either Good Videos or Manual Edit Required. The model demonstrated strong performance, underscoring its robustness and practical applicability in automating video quality evaluation (Table 5.10 and Table 5.11).

Table 5.10 Confusion matrix of extract videos without zoom for validation dataset

	Predicted "Good videos"	**Predicted "Manual edit required videos"**
Actual "Good videos"	23	1
Actual "Manual edit required videos"	1	19

Table 5.11 Performance metrics for extracted videos without zoom for validation dataset

Accuracy	Precision	Sensitivity (Recall)	F1 Score	Specificity	MCC
0.95	0.96	0.96	0.96	0.95	0.91

Out of the videos labeled as good, 23 were correctly classified, while one was incorrectly classified to require manual edits. On the other hand, out of the videos that needed manual edits, 19 were correctly identified with only one misclassified as good. This, therefore, gave an overall accuracy rate of 95%. Precision and recall of the validation dataset were both 0.958, meaning that it is equally good at reducing false positives and false negatives. The F1-score, accordingly, was also 0.958. These results show that the model can maintain strong performance when evaluated on unseen data.

6 Discussion

6.1 Multiple persons detection

The algorithm for the detecting multiple people was successful for all but 3 videos, thereby demonstrating its effectiveness. Accuracy was 95%, and this high accuracy was further supported by 0.97specificity. This demonstrated that the algorithm minimizes false-positive instances. These metrics indicate a well-calibrated system capable of maintaining consistency across diverse inputs.

For the video with only one person classified as multiple people, the MoveNet pose estimation model output a high confidence value for two bounding box detections for the same person.

False negatives occurred when a second person briefly moved into the image. The algorithm failed to find this presence because of low confidence scores, which sheds light on the challenge involved with identifying transient people in a dynamic video environment. It is noted that the person of interest was correctly identified in these frames (i.e., no effect on outcomes).

6.2 Multiple persons overlap detection

The algorithm achieved a sensitivity of 94% and a specificity of 95%, indicating its effectiveness in detecting overlapping. Moreover, the Matthews Correlation Coefficient (MCC) of 0.87 emphasizes its ability to deal with imbalanced datasets, where overlapping situations are much less common than non-overlapping frames.

A detailed analysis of the mistakes showed two major cases where the algorithm struggled. In the first case, the same frame as discussed in section 6.1 caused errors in overlap detection.

A false negative was made during an actual overlap scenario, where the confidence scores assigned to the second person were low and the overlap with the person of interest was very short. As pose detection models mature, multiple person detection errors should reduce.

6.3 Plane detection

The algorithm correctly classified all videos as sagittal, coronal, or transverse. The inclusion of challenging edge cases, such as videos containing zoom events and multiple people, confirmed the algorithm’s robustness.

6.4 Zoom detection

The algorithm caught almost all zoom events. Moreover, algorithm had few false positives,

ensuring that non-zoom events were not misclassified. The F1-score of 0.93 emphasised the balance between precision and recall, further ascertaining the effectiveness of the algorithm in the reduction of false positives and false negatives.

Of the errors that were found, false positives were found in videos labelled as non-zoom, where small variations in bounding box size led to misclassification. These were attributed to factors such as limping that caused changes in the person's estimated height and small variations in the keypoints. Similarly, false negatives were recorded in videos where there were zoom events. In both cases, the zooming action was gradual and subtle, causing changes in bounding box size that fell below the algorithm's detection threshold. These false negatives highlight the difficulty of the algorithm in recognizing small zoom events; however, changes this small should have a minimal effect on automated EVGS scoring.

6.5 Extract video segment without zoom

The algorithm achieved 92% accuracy in the validation dataset, demonstrating that the algorithm can effectively identify video segments longer than 3 seconds and without zoom events.

The errors that occurred could be attributed to cases where the person walked towards the camera when zooming-in, and for the person walking away from camera and zooming-out, especially when the zoom event was subtle. For example, when the change in a person's height due to zooming is smaller than the change in height caused by the person walking toward or away from the camera. Other than this subtle zoom, the algorithm can detect sharp zoom events (sudden change in person's height) effectively.

6.6 Overall video quality classification

The Random Forest classifier achieved high performance when providing an overall gait video quality classification. The system's success can be attributed to the use of multiple quality metrics that enabled a thorough evaluation of the video characteristics. In addition, the probability-based scoring scheme made the classification more interpretable, hence offering clinicians an objective measure of the general quality of the video. Both the overall classification and score (0-100) would be provided to the user.

Two videos were misclassified (videos that needed hand-editing were wrongly determined to be acceptable). This occurred when videos barely passed thresholds for individual criteria. Also, false negatives emerged where slight deviations in one or more metrics resulted in the

misclassification of good videos as needing manual edits. These borderline cases suggest areas where improvement is possible.

The automated video quality assessment method in this thesis could dramatically reduce manual labor in pre-processing gait videos because the integration of the different metrics into one framework makes model-based evaluations uniform and objective. Of course, many opportunities are still open, such as adding complementary features such as keeping track of motion stability and lighting consistency, could help handle marginal instances more robustly. Refining thresholds for individual metrics, based on deeper analysis of misclassified videos, could further reduce these error rates.

7 Summary and Future works

7.1 Research summary

This thesis presents a comprehensive framework for automated video quality assessment in gait analysis, addressing critical challenges in the field of clinical gait evaluation. The research focuses on developing robust algorithms for detecting multiple persons, plane orientation, zoom artifacts, and overall video quality suitable for automated EVGS scoring.

The thesis begins by highlighting the importance of gait analysis in clinical settings and the challenges associated with manual EVGS scoring. Traditional methods are time-consuming and subject to human error, necessitating the development of automated systems. However, the effectiveness of these automated systems heavily depends on the quality of input videos, which led to the primary objective of this research: creating a robust, automated system for assessing the quality of gait analysis videos.

The methodology employed in this research utilizes advanced pose estimation techniques, specifically the MoveNet Multipose Lightning model, to extract skeletal keypoints from video frames. This approach allows for efficient detection of multiple persons and accurate tracking of the patient of interest. The research then developed algorithms for plane detection, multiple person detection, overlap assessment, zoom detection, and resolution evaluation.

The plane detection algorithm demonstrated exceptional accuracy in classifying videos into sagittal, coronal, and transverse views, achieving perfect classification. This high level of accuracy ensures that videos can be appropriately classified by plane of movement, assuming that appropriate guidelines for sagittal and coronal plane video capture are followed (i.e., camera should be in the appropriate orientation for the plane being captured).

Multiple person detection and overlap assessment algorithms showed strong performance, with accuracy rates of over 95%. These algorithms effectively identified scenarios where additional individuals may interfere with the gait analysis, ensuring the focus remains on the person of interest.

The zoom detection algorithm had robust performance, achieving 92% accuracy. This algorithm successfully identified sudden changes in the patient's bounding box height, which may signify unintentional zooming and affect the consistency of gait measurements.

The final video quality assessment framework integrates these individual components into a unified quality scoring system. Utilizing a Random Forest classifier, the system achieved 95%

accuracy in categorizing videos as either suitable for immediate analysis or requiring manual editing. This high level of accuracy demonstrated the effectiveness of the proposed methodology in automating the quality assessment process.

The research also addressed the critical need for actionable feedback in improving video quality. Specific suggestions on improving video quality can be provided to the user when videos fail to meet quality standards, enabling users to rectify issues and enhance the overall efficiency of the gait analysis workflow.

In conclusion, this thesis makes contributions to the field of automated gait analysis by addressing the crucial aspect of video quality assessment. The developed framework demonstrates high outcome measures across various quality metrics, potentially revolutionizing the efficiency of automated EVGS scoring in clinical practice.

7.2 Assumptions and Limitations

7.2.1 Assumptions

The proposed automated video quality assessment framework for gait analysis is based on several key assumptions:

- Good Lighting Conditions: The framework assumes that all gait videos are recorded under adequate lighting conditions, which is essential for accurate pose estimation and keypoint detection.
- Absence of Motion Blur: The system assumes that the recorded videos are free from motion blur, as clear, sharp images are essential for accurate pose estimation and keypoint detection.
- Stable Camera: No substantial camera shake or instability during the recording, ensuring that changes in the person’s position and size are due to their movement rather than camera motion.
- Camera Positioning: Camera is positioned at hip height for both coronal and sagittal plane recordings, ensuring consistent perspective and scale across different videos.
- Limited Plane Orientations: Only three plane orientations are expected in the gait videos: coronal, sagittal, and transverse. This limitation allows for a focused approach in plane detection and classification algorithms.

- Linear Walking Path: The framework assumes that the patient walks in a straight line during the recording, simplifying the gait analysis process and allowing for more consistent measurements across frames.

These assumptions provide a controlled environment for the development and validation of the automated video quality assessment framework, allowing for a focused evaluation of the core methodologies developed in this thesis. In practice, the person recording the video would have considered these guidelines before starting patient video capture.

7.2.2 Limitations

The primary limitation arises from the dataset used for development and validation, which was sourced from a single laboratory (Dataset details discussed in section 3.1). The quality metrics and algorithm design were based on this specific dataset, potentially limiting the generalizability of the approach. The identification of quality metrics was driven by analyzing cases where the automated EVGS algorithm failed, which may not encompass all possible issues in diverse real-world scenarios. Consequently, the automated quality check algorithm should be evaluated with videos from different datasets or clinical environments.

The intended deployment of this system in clinical settings, particularly through a smartphone app for clinical use, introduces additional challenges. Variations in lighting conditions and phone orientations, which were not evaluated in this study, could affect algorithm performance. The current plane detection algorithm is limited to identifying coronal, sagittal, and transverse planes. Videos captured from other angles or orientations may lead to incorrect plane classification and subsequently inaccurate quality scores.

7.3 Future work

To address the current limitation in datasets, a next step is to fine-tune the algorithm and incorporate additional metrics using diverse datasets. This approach could increase system generalizability, ensuring its effectiveness across a wider range of clinical environments and patient populations. Incorporating a lighting quality metric in future development would be valuable since poor lighting can adversely affect pose estimation model performance [126], which in turn affects gait analysis. Implementing standard image processing techniques, such as analyzing the mean pixel value of video frames, can provide valuable insights into lighting quality.

The current algorithm's ability to track the person of interest in multi-person scenarios can be leveraged to automatically crop videos, using bounding box coordinates of person of interest. Furthermore, identifying frames with specific flaws, such as zoom events or multiple persons, opens the possibility of automatically trimming videos to retain only the high-quality segments. These enhancements could potentially transform videos that currently require manual editing into suitable inputs for automated EVGS algorithms, streamlining the workflow in clinical settings.

While the current implementation utilizes MoveNet Lightning Multipose model for its balance of speed and accuracy, future technological advancements in pose estimation should be closely monitored. As more accurate and lightweight models emerge, replacing MoveNet with a superior alternative could substantially improve the precision of metrics scores and overall effectiveness of the quality assessment algorithm. Additionally, future research could explore the integration of machine learning techniques to refine the classification of video quality. By training models on larger, more diverse datasets, we could potentially identify subtle quality indicators that are not captured by the current rule-based system.

References

[1] M. Whittle, *Gait analysis: an introduction*, 4th ed. Edinburgh ; Butterworth-Heinemann, 2007.

[2] A. Middleton and S. L. Fritz, "Assessment of Gait, Balance, and Mobility in Older Adults: Considerations for Clinicians," *Current Translational Geriatrics and Experimental Gerontology Reports*, vol. 2, no. 4, Art. no. 4, Aug. 2013, doi: 10.1007/s13670-013-0057-2.

[3] S. H. Ramesh, E. D. Lemaire, A. Tu, K. Cheung, and N. Baddour, "Automated Implementation of the Edinburgh Visual Gait Score (EVGS) Using OpenPose and Handheld Smartphone Video," *Sensors (Basel, Switzerland)*, vol. 23, no. 10, pp. 4839-, 2023, doi: 10.3390/s23104839.

[4] R. Rodrigues *et al.*, "Objective quality assessment of medical images and videos: review and challenges," *Multimedia Tools and Applications*, pp. 1–34, Oct. 2024, doi: 10.1007/s11042-024-20292-x.

[5] M. Razaak and M. G. Martini, "Medical image and video quality assessment in e-health applications and services," IEEE, 2013, pp. 6–10. doi: 10.1109/HealthCom.2013.6720628.

[6] M. Leszczuk, "Assessing Task-Based Video Quality — A Journey from Subjective Psycho-Physical Experiments to Objective Quality Models," doi: 10.1007/978-3-642-21512-4_11.

[7] M. Leszczuk, "Optimising task-based video quality," *Multimedia Tools and Applications*, vol. 68, no. 1, Art. no. 1, Jul. 2012, doi: 10.1007/s11042-012-1161-6.

[8] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[9] M. Mundt *et al.*, "Automating Video-Based Two-Dimensional Motion Analysis in Sport? Implications for Gait Event Detection, Pose Estimation, and Performance Parameter Analysis," *Scandinavian journal of medicine & science in sports*, vol. 34, no. 7, pp. e14693-n/a, 2024, doi: 10.1111/sms.14693.

[10] J. M. Irvine and R. J. Wood, "Context and quality estimation in video for enhanced event detection," SPIE, 2015, pp. 94600L-94600L–8. doi: 10.1117/12.2177155.

[11] R. A. States, J. J. Krzak, Y. Salem, E. M. Godwin, A. W. Bodkin, and M. L. McMulkin,

"Instrumented gait analysis for management of gait disorders in children with cerebral palsy: A scoping review," *Gait & Posture*, vol. 90, pp. 1–8, Oct. 2021, doi: 10.1016/j.gaitpost.2021.07.009.

[12] R. Baker, A. Esquenazi, M. G. Benedetti, and K. Desloovere, "Gait analysis: clinical facts".

[13] A. M. L. Ong, S. J. Hillman, and J. E. Robb, "Reliability and validity of the Edinburgh Visual Gait Score for cerebral palsy when used by inexperienced observers," *Gait & Posture*, vol. 28, no. 2, pp. 323–326, Aug. 2008, doi: 10.1016/j.gaitpost.2008.01.008.

[14] M. del Pilar Duque Orozco *et al.*, "Reliability and validity of Edinburgh visual gait score as an evaluation tool for children with cerebral palsy," *Gait & Posture*, vol. 49, pp. 14–18, Sep. 2016, doi: 10.1016/j.gaitpost.2016.06.017.

[15] V. A. Kulkarni, D. T. Kephart, R. Olleac, and J. R. Davids, "Enhancing Observational Gait Analysis – Techniques and Tips for Analyzing Gait Without a Gait Lab," *Journal of the Pediatric Orthopaedic Society of North America*, vol. 2, no. 3, p. 135, Nov. 2020, doi: 10.55275/JPOSNA-2020-135.

[16] J. Robinson, J. Dixon, A. Macsween, P. van Schaik, and D. Martin, "The effects of exergaming on balance, gait, technology acceptance and flow experience in people with multiple sclerosis: a randomized controlled trial," *BMC Sports Science, Medicine and Rehabilitation*, vol. 7, no. 1, p. 8, Apr. 2015, doi: 10.1186/s13102-015-0001-1.

[17] J. Adolf, J. Dolezal, P. Kutilek, J. Hejda, and L. Lhotska, "Single Camera-Based Remote Physical Therapy: Verification on a Large Video Dataset," *Applied Sciences*, vol. 12, no. 2, Art. no. 2, Jan. 2022, doi: 10.3390/app12020799.

[18] J. Zhu, A. Ak, C. Dormeval, P. Le Callet, K. Rahul, and S. Sethuraman, "Subjective test environments: A multifaceted examination of their impact on test results," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, in IMX '23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 298–302. doi: 10.1145/3573381.3596470.

[19] D. K. Dewangan and Y. Rathore, "Image Quality estimation of Images using Full Reference and No Reference Method," *International Journal of Advanced Research in Computer Science*, vol. 2, no. 5, Art. no. 5, Jan. 2017, doi: 10.26483/ijarcs.v2i5.796.

[20] S. Winkler, "Perceptual Video Quality Metrics — A Review," in *Digital Video Image Quality and Perceptual Coding*, CRC Press, 2006.

[21] S. Sonawane and A. M. Deshpande, “Image Quality Assessment Techniques: An Overview,” *International Journal of Engineering Research*, vol. 3, no. 4, 2014.

[22] V. K. Bhola, T. Sharma, and J. Bhatnagar, “Image Quality Assessment Techniques,” 2014.

[23] A. Chetouani, “Full reference image quality assessment: limitation,” in *2014 22nd International Conference on Pattern Recognition*, Aug. 2014, pp. 833–837. doi: 10.1109/ICPR.2014.153.

[24] P. Ndajah, H. Kikuchi, M. Yukawa, H. Watanabe, and S. Muramatsu, “An Investigation on The Quality of Denoised Images,” *International Journal of Circuits, Systems and Signal Processing*.

[25] P. Ndajah, H. Kikuchi, M. Yukawa, H. Watanabe, and S. Muramatsu, “SSIM Image Quality Metric for Denoised Images”.

[26] M. G. Martini, B. Villarini, and F. Fiorucci, “A reduced-reference perceptual image and video quality metric based on edge preservation,” *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, Art. no. 1, Mar. 2012, doi: 10.1186/1687-6180-2012-66.

[27] M. Nuutinen, O. Orenius, T. Säämänen, and P. Oittinen, “Reference image method for measuring quality of photographs produced by digital cameras,” Accessed: Feb. 06, 2025. [Online]. Available: https://www-spiedigitallibrary-org.proxy.bib.uottawa.ca/conference-proceedings-of-spie/7867/78670M/Reference-image-method-for-measuring-quality-of-photographs-produced-by/10.1117/12.871999.full

[28] S. Dost, F. Saud, M. Shabbir, M. G. Khan, M. Shahid, and B. Lovstrom, “Reduced reference image and video quality assessments: review of methods,” *J Image Video Proc.*, vol. 2022, no. 1, p. 1, Jan. 2022, doi: 10.1186/s13640-021-00578-y.

[29] M. Shahid, K. Pandremmenou, L. P. Kondi, A. Rossholm, and B. Lövström, “Perceptual quality estimation of H.264/AVC videos using reduced-reference and no-reference models,” *J. Electron. Imaging*, vol. 25, no. 5, p. 053012, Sep. 2016, doi: 10.1117/1.JEI.25.5.053012.

[30] W. Zhou, G. Yue, R. Zhang, Y. Qin, and H. Liu, “Reduced-Reference Quality Assessment of Point Clouds via Content-Oriented Saliency Projection,” *IEEE Signal Process. Lett.*, vol. 30, pp. 354–358, 2023, doi: 10.1109/LSP.2023.3264105.

[31] H. Otroshi Shahreza, A. Amini, and H. Behroozi, “In-the-wild No-Reference image quality assessment using Deep Convolutional Neural Networks,” in *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, Dec. 2019, pp. 1–4. doi:

10.1109/ICSPIS48872.2019.9066036.

[32] D. Varga, “No-Reference Image Quality Assessment with Convolutional Neural Networks and Decision Fusion,” *Applied Sciences*, vol. 12, no. 1, p. 101, Dec. 2021, doi: 10.3390/app12010101.

[33] Y. Li, L.-M. Po, L. Feng, and F. Yuan, “No-reference image quality assessment with deep convolutional neural networks,” in *2016 IEEE International Conference on Digital Signal Processing (DSP)*, Oct. 2016, pp. 685–689. doi: 10.1109/ICDSP.2016.7868646.

[34] L. Wang, “A survey on IQA,” Jan. 11, 2022, *arXiv*: arXiv:2109.00347. doi: 10.48550/arXiv.2109.00347.

[35] A. De Decker, J. De Cock, and G. Van Wallendael, “Combining deep learning and feature engineering for No-Reference video quality assessment,” *Proceedings of the 3rd Mile-High Video Conference on zzz*, pp. 120–121, Feb. 2024, doi: 10.1145/3638036.3640287.

[36] M. H. Pinson, “Why No Reference Metrics for Image and Video Quality Lack Accuracy and Reproducibility,” *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 97–117, Mar. 2023, doi: 10.1109/TBC.2022.3191059.

[37] P. Ye and D. Doermann, “No-Reference Image Quality Assessment Using Visual Codebooks,” *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, Jul. 2012, doi: 10.1109/TIP.2012.2190086.

[38] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, “Generalizable No-Reference Image Quality Assessment via Deep Meta-Learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1048–1060, Mar. 2022, doi: 10.1109/TCSVT.2021.3073410.

[39] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, “UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content,” *IEEE Trans. on Image Process.*, vol. 30, pp. 4449–4464, 2021, doi: 10.1109/TIP.2021.3072221.

[40] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘Completely Blind’ Image Quality Analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013, doi: 10.1109/LSP.2012.2227726.

[41] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using swin transformer,” in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844. doi: 10.1109/ICCVW54120.2021.00210.

[42] T. Wiegand, “Draft itu-t recommendation and final draft international standard on joint video specification,” *H. 264/ISO/IEC 14496-10 AVC, JVT-G050*, 2003, Accessed: Feb. 06, 2025. [Online]. Available: https://cir.nii.ac.jp/crid/1572824500866259584

[43] H. Wu *et al.*, “DisCoVQA: Temporal Distortion-Content Transformers for Video Quality Assessment,” *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 33, no. 9, pp. 4840–4854, Sep. 2023, doi: 10.1109/TCSVT.2023.3249741.

[44] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2408–2415. doi: 10.1109/CVPR.2012.6247954.

[45] H. Wu *et al.*, “Exploring video quality assessment on user generated contents from aesthetic and technical perspectives,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 20087–20097. doi: 10.1109/ICCV51070.2023.01843.

[46] M. Mirkovic, P. Vrgovic, D. Culibrk, D. Stefanovic, and A. Anderla, “Evaluating the Role of Content in Subjective Video Quality Assessment,” *The Scientific World Journal*, vol. 2014, no. 1, p. 625219, 2014, doi: 10.1155/2014/625219.

[47] B. Furht, Ed., *Encyclopedia of Multimedia*, 1st edition. New York: Springer, 2005.

[48] C. Delgorge, C. Rosenberger, G. Poisson, and P. Vieyres, “Towards a New Tool for the Evaluation of the Quality of Ultrasound Compressed Images,” *IEEE Transactions on Medical Imaging*, vol. 25, no. 11, pp. 1502–1509, Nov. 2006, doi: 10.1109/TMI.2006.883088.

[49] L. Soares, L. Borges, B. Barufaldi, A. Maidment, and M. A. Vieira, “Using virtual clinical trials to assess objective image quality metrics in the task of microcalcification localization in digital mammography,” in *16th International Workshop on Breast Imaging (IWBI2022)*, H. Bosmans, N. Marshall, and C. Van Ongeval, Eds., Leuven, Belgium: SPIE, Jul. 2022, p. 34. doi: 10.1117/12.2625745.

[50] L. S. Chow, H. Rajagopal, and R. Paramesran, “Correlation between subjective and objective assessment of magnetic resonance (MR) images,” *Magnetic Resonance Imaging*, vol. 34, no. 6, pp. 820–831, Jul. 2016, doi: 10.1016/j.mri.2016.03.006.

[51] I. A. Kowalik-Urbaniak *et al.*, “Modelling of Subjective Radiological Assessments with Objective Image Quality Measures of Brain and Body CT Images,” Jan. 2015. doi:

10.1007/978-3-319-20801-5_1.

[52] A. Panayides, Z. C. Antoniou, Y. Mylonas, M. S. Pattichis, A. Pitsillides, and C. S. Pattichis, "High-Resolution, Low-Delay, and Error-Resilient Medical Ultrasound Video Communication Using H.264/AVC Over Mobile WiMAX Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 619–628, May 2013, doi: 10.1109/TITB.2012.2232675.

[53] M. Razaak, M. G. Martini, and K. Savino, "A Study on Quality Assessment for Medical Ultrasound Video Compressed via HEVC," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1552–1559, Sep. 2014, doi: 10.1109/JBHI.2014.2326891.

[54] Z. A. Khan, A. Beghdadi, M. Kaaniche, F. Alaya-Cheikh, and O. Gharbi, "A neural network based framework for effective laparoscopic video quality assessment," *Computerized Medical Imaging and Graphics*, vol. 101, p. 102121, Oct. 2022, doi: 10.1016/j.compmedimag.2022.102121.

[55] M. A. Usman *et al.*, "Suitability of VVC and HEVC for Video Telehealth Systems," Jan. 2021, doi: 10.32604/cmc.2021.014614.

[56] A. Sekhri, S. A. Amirshahi, and M.-C. Larabi, "Enhancing Content Representation for AR Image Quality Assessment Using Knowledge Distillation," Dec. 08, 2024, *arXiv*: arXiv:2412.06003. doi: 10.48550/arXiv.2412.06003.

[57] S. E. Ghazouali, U. Michelucci, Y. E. Hillali, and H. Nouira, "CSIM: A Copula-based similarity index sensitive to local changes for Image quality assessment," Oct. 04, 2024, *arXiv*: arXiv:2410.01411. doi: 10.48550/arXiv.2410.01411.

[58] H. Kaur, R. Vig, N. Kumar, A. Dogra, A. Sharma, and B. Goyal, "Objective image quality assessment of pixel level image fusion algorithms for medical imaging," *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 01–08, Apr. 2023, doi: 10.1109/ICEEICT56924.2023.10157703.

[59] Dr. Smita Nirkhi *et al.*, "GANs in medical imaging: synthesizing of realistic images for analysis," *IJARSCT*, pp. 415–420, May 2024, doi: 10.48175/IJARSCT-18557.

[60] L. Lévêque *et al.*, "On the Subjective Assessment of the Perceived Quality of Medical Images and Videos," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2018, pp. 1–6. doi: 10.1109/QoMEX.2018.8463297.

[61] L. Lévêque, M. Outtas, H. Liu, and L. Zhang, “Comparative study of the methodologies used for subjective medical image quality assessment,” *Phys. Med. Biol.*, vol. 66, no. 15, p. 15TR02, Jul. 2021, doi: 10.1088/1361-6560/ac1157.

[62] Y. Ding, “Medical Image Quality Assessment,” 2018. doi: 10.1007/978-3-662-56497-4_8.

[63] J. Shiraishi, L. L. Pesce, C. E. Metz, and K. Doi, “Experimental Design and Data Analysis in Receiver Operating Characteristic Studies: Lessons Learned from Reports in Radiology from 1997 to 2006,” *Radiology*, vol. 253, no. 3, pp. 822–830, Dec. 2009, doi: 10.1148/radiol.2533081632.

[64] B. Kim, K. H. Lee, K. J. Kim, R. Mantiuk, H. Kim, and Y. H. Kim, “Artifacts in Slab Average-Intensity-Projection Images Reformatted from JPEG 2000 Compressed Thin-Section Abdominal CT Data Sets,” *American Journal of Roentgenology*, vol. 190, no. 6, pp. W342–W350, Jun. 2008, doi: 10.2214/AJR.07.3405.

[65] O. Meriem, Z. Lu, and M. Maria, “Towards Recommendations and Guidelines for Subjective Medical Image and Video Quality Assessment,” in *2024 12th European Workshop on Visual Information Processing (EUVIP)*, Sep. 2024, pp. 1–6. doi: 10.1109/EUVIP61797.2024.10772875.

[66] C. Delgorge, C. Rosenberger, G. Poisson, and P. Vieyres, “Towards a New Tool for the Evaluation of the Quality of Ultrasound Compressed Images,” *IEEE Transactions on Medical Imaging*, vol. 25, no. 11, pp. 1502–1509, Nov. 2006, doi: 10.1109/TMI.2006.883088.

[67] A. B. Mansoor, M. Haider, A. S. Mian, and S. A. Khan, “A Hybrid Image Quality Measure for Automatic Image Quality Assessment,” doi: 10.1007/978-3-642-02230-2_10.

[68] C. Cavaro-Menard, L. Zhang, and P. Le Callet, “Diagnostic quality assessment of medical images: Challenges and trends,” in *2010 2nd European Workshop on Visual Information Processing (EUVIP)*, Jul. 2010, pp. 277–284. doi: 10.1109/EUVIP.2010.5699147.

[69] N. Sinha and A. G. Ramakrishnan, “Quality Assessment in Magnetic Resonance Images”, Accessed: Jan. 15, 2025. [Online]. Available: https://www.dl.begellhouse.com/journals/4b27cbfc562e21b8,4af52eb12a5f0d0c,6e3f59670d54c651.html

[70] M. Perez-Diaz, “Techniques to evaluate the quality of medical images,” *AIP Conference Proceedings*, vol. 1626, no. 1, pp. 39–45, Nov. 2014, doi: 10.1063/1.4901358.

[71] J.-W. Oestmann and M. Galanski, “ROC: Methodik zum Vergleich der diagnostischen Leistung bildgebender Verfahren,” *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 151, pp. 89–92, Mar. 2008, doi: 10.1055/s-2008-1047135.

[72] C. E. Metz, “Fundamental ROC Analysis”, Accessed: Jan. 15, 2025. [Online]. Available: https://www.spiedigitallibrary.org/ebooks/PM/Handbook-of-Medical-Imaging-Volume-1-Physics-and-Psychophysics/15/Fundamental-ROC-Analysis/10.1117/3.832716.ch15

[73] L. S. Chow, H. Rajagopal, and R. Paramesran, “Correlation between subjective and objective assessment of magnetic resonance (MR) images,” *Magnetic Resonance Imaging*, vol. 34, no. 6, pp. 820–831, Jul. 2016, doi: 10.1016/j.mri.2016.03.006.

[74] I. A. Kowalik-Urbaniak *et al.*, “Modelling of Subjective Radiological Assessments with Objective Image Quality Measures of Brain and Body CT Images,” Jan. 2015. doi: 10.1007/978-3-319-20801-5_1.

[75] K. Ohashi *et al.*, “Applicability Evaluation of Full-Reference Image Quality Assessment Methods for Computed Tomography Images,” *Journal of Digital Imaging*, vol. 36, no. 6, Art. no. 6, Aug. 2023, doi: 10.1007/s10278-023-00875-0.

[76] A. Mason *et al.*, “Comparison of Objective Image Quality Metrics to Expert Radiologists’ Scoring of Diagnostic Quality of MR Images,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1064–1072, Apr. 2020, doi: 10.1109/TMI.2019.2930338.

[77] G. Ginesu, F. Massidda, and D. D. Giusto, “A multi-factors approach for image quality assessment based on a human visual system model,” *Signal Processing: Image Communication*, vol. 21, no. 4, pp. 316–333, Apr. 2006, doi: 10.1016/j.image.2005.11.005.

[78] F. Zhang and D. R. Bull, “A Perception-Based Hybrid Model for Video Quality Assessment,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, Jun. 2016, doi: 10.1109/TCSVT.2015.2428551.

[79] H. Liu and I. Heynderickx, “Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, Jul. 2011, doi: 10.1109/TCSVT.2011.2133770.

[80] L. Zhang, C. Cavaro-Ménard, and P. L. Callet, “Key issues and specificities for the objective medical image quality assessment”.

[81] L. S. Chow and R. Paramesran, “Review of medical image quality assessment,” *Biomedical*

Signal Processing and Control, vol. 27, pp. 145–154, May 2016, doi: 10.1016/j.bspc.2016.02.006.

[82] B. L. Eck *et al.*, "Computational and human observer image quality evaluation of low dose, knowledge-based CT iterative reconstruction," *Medical Physics*, vol. 42, no. 10, pp. 6098–6111, 2015, doi: 10.1118/1.4929973.

[83] L. Zhang, C. Cavaro-Ménard, and P. Le Callet, "An overview of model observers," *IRBM*, vol. 35, no. 4, pp. 214–224, Sep. 2014, doi: 10.1016/j.irbm.2014.04.002.

[84] J. G. Brankov, "Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection," *Phys. Med. Biol.*, vol. 58, no. 20, p. 7159, Sep. 2013, doi: 10.1088/0031-9155/58/20/7159.

[85] J. Greffier, J. Frandon, A. Larbi, J. P. Beregi, and F. Pereira, "CT iterative reconstruction algorithms: a task-based image quality assessment," *European Radiology*, vol. 30, no. 1, Art. no. 1, Jul. 2019, doi: 10.1007/s00330-019-06359-6.

[86] K. Li, W. Zhou, H. Li, and M. A. Anastasio, "Assessing the Impact of Deep Neural Network-Based Image Denoising on Binary Signal Detection Tasks," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2295–2305, Sep. 2021, doi: 10.1109/TMI.2021.3076810.

[87] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 21, pp. 9758–9765, Nov. 1993, doi: 10.1073/pnas.90.21.9758.

[88] M. Alnowami *et al.*, "A deep learning model observer for use in alterative forced choice virtual clinical trials," Accessed: Jan. 15, 2025. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10577/105770Q/A-deep-learning-model-observer-for-use-in-alterative-forced/10.1117/12.2293209.full

[89] F. Massanes and J. G. Brankov, "Evaluation of CNN as anthropomorphic model observer," Accessed: Jan. 15, 2025. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10136/101360Q/Evaluation-of-CNN-as-anthropomorphic-model-observer/10.1117/12.2254603.full

[90] H.-W. Tseng, J. Fan, and M. A. Kupinski, "Design of a practical model-observer-based image quality assessment method for x-ray computed tomography imaging systems", Accessed: Jan. 15, 2025. [Online]. Available:

https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-3/issue-3/035503/Design-of-a-practical-model-observer-based-image-quality-assessment/10.1117/1.JMI.3.3.035503.full

[91] L. Wu, J.-Z. Cheng, S. Li, B. Lei, T. Wang, and D. Ni, "FUIQA: Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks," *IEEE Transactions on Cybernetics*, vol. 47, no. 5, pp. 1336–1349, May 2017, doi: 10.1109/TCYB.2017.2671898.

[92] W. Zhou, H. Li, and M. A. Anastasio, "Approximating the Ideal Observer for Joint Signal Detection and Localization Tasks by use of Supervised Learning Methods," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3992–4000, Dec. 2020, doi: 10.1109/TMI.2020.3009022.

[93] I. Lorente, C. K. Abbey, and J. G. Brankov, "Deep learning based model observer by U-Net," Accessed: Jan. 15, 2025. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11316/113160F/Deep-learning-based-model-observer-by-U-Net/10.1117/12.2549687.full

[94] C. Cavaro-Menard, L. Zhang, and P. Le Callet, "Diagnostic quality assessment of medical images: Challenges and trends," in *2010 2nd European Workshop on Visual Information Processing (EUVIP)*, Jul. 2010, pp. 277–284. doi: 10.1109/EUVIP.2010.5699147.

[95] H. Lin and H. Gifford, "Addition of a threshold mechanism to model observers for medical image quality assessment," in *Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment*, Y. Chen and C. R. Mello-Thoms, Eds., San Diego, United States: SPIE, Mar. 2024, p. 14. doi: 10.1117/12.3008279.

[96] F. Valeri *et al.*, "UNet and MobileNet CNN-based model observers for CT protocol optimization: comparative performance evaluation by means of phantom CT images," *J Med Imaging (Bellingham)*, vol. 10, no. Suppl 1, p. S11904, Feb. 2023, doi: 10.1117/1.JMI.10.S1.S11904.

[97] R. A. Welikala *et al.*, "Automated retinal image quality assessment on the UK Biobank dataset for epidemiological studies," *Computers in Biology and Medicine*, vol. 71, pp. 67–76, Apr. 2016, doi: 10.1016/j.compbiomed.2016.01.027.

[98] R. Rodrigues and A. M. G. Pinheiro, "Segmentation of Skeletal Muscle in Thigh Dixon MRI Based on Texture Analysis," Apr. 09, 2019, *arXiv*: arXiv:1904.04747. doi: 10.48550/arXiv.1904.04747.

[99] R. Alais, P. Dokládal, A. Erginay, B. Figliuzzi, and E. Decencière, “Fast macula detection and application to retinal image quality assessment,” *Biomedical Signal Processing and Control*, vol. 55, p. 101567, Jan. 2020, doi: 10.1016/j.bspc.2019.101567.

[100] J. Hu, C. Zhang, K. Zhou, and S. Gao, “Chest X-Ray Diagnostic Quality Assessment: How Much Is Pixel-Wise Supervision Needed?,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 7, pp. 1711–1723, Jul. 2022, doi: 10.1109/TMI.2022.3149171.

[101] A. Reinke *et al.*, “Common Limitations of Image Processing Metrics: A Picture Story,” Dec. 06, 2023, *arXiv*: arXiv:2104.05642. doi: 10.48550/arXiv.2104.05642.

[102] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, “Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation,” *Medical Image Analysis*, vol. 63, p. 101693, Jul. 2020, doi: 10.1016/j.media.2020.101693.

[103] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, “2D Human pose estimation: a survey,” *Multimedia Syst.*, vol. 29, no. 5, pp. 3115–3138, Nov. 2022, doi: 10.1007/s00530-022-01019-0.

[104] Y. Hou, J. Li, S. Liao, and N. Xue, “Research Advanced in Human Pose Estimation based on Deep Learning,” vol. 119, 2024.

[105] G. Ning, P. Liu, X. Fan, and C. Zhang, “A Top-Down Approach to Articulated Human Pose Estimation and Tracking,” Jan. 2019. doi: 10.1007/978-3-030-11012-3_20.

[106] T. D. Nguyen and M. Kresovic, “A survey of top-down approaches for human pose estimation,” Feb. 05, 2022, *arXiv*: arXiv:2202.02656. doi: 10.48550/arXiv.2202.02656.

[107] M. Li, Z. Zhou, J. Li, and X. Liu, “Bottom-up Pose Estimation of Multiple Person with Bounding Box Constraint,” Jul. 26, 2018, *arXiv*: arXiv:1807.09972. doi: 10.48550/arXiv.1807.09972.

[108] Y. Guo, J. Liu, G. Li, L. Mai, and H. Dong, “Fast and flexible human pose estimation with HyperPose,” in *Proceedings of the 29th ACM International Conference on Multimedia*, Oct. 2021, pp. 3763–3766. doi: 10.1145/3474085.3478325.

[109] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” May 30, 2019, *arXiv*: arXiv:1812.08008. doi: 10.48550/arXiv.1812.08008.

[110] A. Mathis *et al.*, “DeepLabCut: markerless pose estimation of user-defined body parts with deep learning,” *Nature Neuroscience*, vol. 21, no. 9, pp. 1281–1289, Sep. 2018, doi:

10.1038/s41593-018-0209-y.

[111] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, “PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model,” Mar. 22, 2018, *arXiv*: arXiv:1803.08225. doi: 10.48550/arXiv.1803.08225.

[112] A. Toshev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1653–1660. doi: 10.1109/CVPR.2014.214.

[113] “Movenet singlepose lightning.” [Online]. Available: https://www.kaggle.com/models/google/movenet/tensorFlow2/singlepose-lightning

[114] “Movenet Thunder.” [Online]. Available: https://www.kaggle.com/models/google/movenet/tensorFlow2/singlepose-thunder

[115] “Movenet multipose lightning.” [Online]. Available: https://www.kaggle.com/models/google/movenet/tensorFlow2/multipose-lightning

[116] “Comparative Analysis of OpenPose, PoseNet, and MoveNet Models for Pose Estimation in Mobile Devices | IIETA.” Accessed: Nov. 16, 2024. [Online]. Available: https://www.iieta.org/journals/ts/paper/10.18280/ts.390111

[117] “COCO - Common Objects in Context.” Accessed: Jan. 16, 2025. [Online]. Available: https://cocodataset.org/#home

[118] “MPII Human Pose Dataset - Max Planck Institute for Informatics.” Accessed: Jan. 16, 2025. [Online]. Available: https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/software-and-datasets/mpii-human-pose-dataset

[119] “MoveNet Architecture.” [Online]. Available: https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html

[120] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” Apr. 19, 2017, *arXiv*: arXiv:1612.03144. doi: 10.48550/arXiv.1612.03144.

[121] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint Triplets for Object Detection,” Apr. 19, 2019, *arXiv*: arXiv:1904.08189. doi: 10.48550/arXiv.1904.08189.

[122] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted

Residuals and Linear Bottlenecks," Mar. 21, 2019, *arXiv*: arXiv:1801.04381. doi: 10.48550/arXiv.1801.04381.

[123] "performance metrics." [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

[124] S. Yadav, "Why is the f1-score the harmonic mean of precision and recall rather than the arithmetic mean?," Medium. Accessed: Jan. 23, 2025. [Online]. Available: https://medium.com/@Suraj_Yadav/why-is-the-f1-score-the-harmonic-mean-of-precision-and-recall-rather-than-the-arithmetic-mean-2573ab99e49c

[125] Anishnama, "Matthews Correlation Coefficient(MCC) one of the best metric when 2 classes are imbalanced," Medium. Accessed: Jan. 23, 2025. [Online]. Available: https://medium.com/@anishnama20/matthews-correlation-coefficient-mcc-one-of-the-best-metric-when-2-classes-are-imbalanced-c0318ac68c21

[126] S. Lee *et al.*, "Human pose estimation in extremely low-light conditions," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 704–714. doi: 10.1109/CVPR52729.2023.00075.