HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

Viability of Automating Annotation Through MediaPipe's Effectiveness on Pose Points Model Accuracy

Dhruv Jena[1], Daksh Jain[1] and Anthony Mauro[#]

[1]Mountain House High School, USA
[#]Advisor

ABSTRACT

This research investigates the effectiveness of automated annotation using MediaPipe for human motion recognition tasks, comparing its performance against manually annotated data from the MP2 dataset. Through the evaluation of three machine learning models—Generative Adversarial Network (GAN), Dense model, and Transformer model—applied to both datasets, we assess the impact of dataset quality and annotation methods on model performance. The findings indicate that while models trained on MediaPipe-annotated data generally outperformed those trained on MP2, the overall accuracy remained low across all models, highlighting challenges in generalization. The study identifies the need for high-quality automated annotations that approach the granularity of manual annotations to improve performance. Moreover, it suggests that environmental factors such as lighting, background, and camera angles, which can affect joint detection accuracy, contribute to performance inconsistencies. The research also emphasizes the importance of data preprocessing, augmentation, and the potential for combining multi-modal data to enhance annotation precision. Ultimately, this study demonstrates that automated annotation offers scalability for large-scale projects but requires refinement in handling complex, dynamic environments to fully realize its potential in machine learning applications.

Introduction

Background and Motivation

Recent fast advances in machine learning and computer vision have led to increasing interest in automated data annotation methods (Quiñonez et al., 2022; Lugaresi et al., 2019). The ability to obtain high-quality annotated data is a fundamental requisite for developing robust machine learning models, especially in applications that require large datasets, such as human activity recognition, autonomous driving, and healthcare diagnosis (Radeta et al., 2024). Traditional manual annotation methods are exhaustive but too time-consuming and costly when applied to large datasets of photos, therefore determining the need for automated annotation solutions.

Labeling can be optimized to use automated annotation with pre-trained models and advanced tools like MediaPipe or Ultralytic's YOLOv8 (Lugaresi et al., 2019; Radeta et al., 2024). MediaPipe is an all-inclusive framework developed by Google for real-time perception models used in tasks like hand tracking, pose estimation, and object detection, enabling real-time annotation with minimal human effort. The integration of these tools is especially useful in domains with immediate responses or those with large amounts of annotation (K et al., 2021). With MediaPipe, it becomes possible for practitioners to use models for initial labeling, which can later be fine-tuned, increasing the efficiency and scalability of data preparation.

Research Objectives

The main goals of this study are to assess the performance of automated annotation with MediaPipe and to explore the following research questions:

1. How well can automated annotation using MediaPipe compare to manual annotation across different photo contexts?
2. How does automated annotation affect data consistency and model performance for large machine learning projects?
3. What specific challenges do automated annotation systems encounter when applied to diverse video domains, and how might these problems be addressed?

By investigating these questions, our study tries to determine the practical feasibility of MediaPipe in the scope of large-scale video annotation while providing information about its ability to address time and economic constraints.

Significance of the Study

The quality of the labeled data has a very high impact on the successes that machine learning (ML) models achieve in real applications. Precise annotation lets models generalize well to ensure reliability when it comes to tasks such as autonomous driving or human activity recognition (Radeta et al., 2024). Given the laborious nature associated with the task of manual annotation for videos, automated annotation methods are crucial for scalability in ML. This study, therefore, explores MediaPipe as an annotation tool that satisfies the demand for a more efficient and accurate labeling solution for ML and contributes valuable knowledge to enhance large-scale data preparation methods.

Literature Review

Advancements in data preparation for machine learning highlight both the challenges of annotation and the potential of emerging tools to improve efficiency and quality.

Challenges in Photo Data Annotation

The process of annotating photo data is crucial in the development of machine learning models for computer vision tasks such as object detection, image segmentation, and facial recognition. However, manual annotation of photo datasets presents several challenges.

One major issue is the sheer volume of data that must be processed. Large-scale datasets often consist of tens of thousands or even millions of images, each requiring detailed and precise annotations. Unlike videos, which require tracking over time, photo datasets often involve annotating highly varied content across a broad set of individual images, each with unique contexts and features. This variety can make the process labor-intensive and time-consuming. Maintaining accuracy and consistency across the dataset is another significant challenge. Human annotators, particularly when working with large datasets under tight deadlines, are prone to errors such as mislabeling or inconsistencies in annotation standards. These errors can degrade the quality of the dataset and, consequently, the performance of machine learning models trained on it. Furthermore, individual interpretations of labeling instructions or subjective differences among annotators may introduce biases or discrepancies, reducing the reliability and robustness of the annotations (Ward et al., 2021).

These challenges have spurred the development of automated and semi-automated annotation techniques to minimize manual effort while improving data quality and consistency.

Overview of MediaPipe

MediaPipe is an open-source framework developed by Google that offers a wide range of pre-trained models and tools specifically designed for real-time computer vision and machine learning applications. It offers a flexible and efficient platform applicable to various tasks: hand tracking, pose estimation, facial landmark detection, and object detection, among others.

Real-time processing is probably one of the most striking characteristics of MediaPipe; hence, it would be well-suited for video annotation tasks requiring fast analysis. MediaPipe uses deep learning algorithms in the detection and tracking of objects, activities, or key points on video frames to provide an automated or semi-automated process for annotation.

MediaPipe has been widely applied in augmented reality, robotics, and video surveillance due to its effectiveness in analyzing video data. In the domain of video data annotation, the object detection and pose estimation models provided by MediaPipe have been used for automating keypoint labeling or object tracking across frames, hence saving much time and labor associated with manual annotation. It is, however, very sensitive to input data quality—issues such as video resolution and lighting conditions will result in a very different accuracy (Zhang et al., 2021).

Overview of ML Models

These machine-learning models all demand different types of data, with various levels of complexity and dependence on annotated data. The next section summarizes the most important models used in the analysis of photo data.

1. Dense Networks: Dense networks, particularly Dense Transformer Networks, integrate convolutional layers with spatial transformer modules to improve pixel-level predictions in tasks like image segmentation. By dynamically learning patch sizes and shapes based on local image statistics, they address the limitations of fixed patch sizes in traditional CNNs. These networks use an encoder-decoder architecture to ensure spatial correspondence between input images and output label maps, enhancing their performance on dense prediction tasks such as object localization and feature extraction (Hochreiter & Schmidhuber, 1997).
2. Generative Adversarial Networks (GANs): GANs are widely used for generating high-fidelity synthetic data, including images and videos. They consist of a generator that creates samples and a discriminator that evaluates their authenticity, pushing the generator to improve. GANs excel in creative and data augmentation applications but require careful tuning to avoid instabilities like mode collapse. They are instrumental in producing data where annotations are sparse, such as in medical imaging or creative design (Sherstinsky, 2020).
3. Transformer Models: Transformers, originally designed for natural language processing, have expanded into computer vision with architectures like the Vision Transformer (ViT). These models rely on self-attention mechanisms to capture long-range dependencies and spatial relationships, outperforming traditional CNNs in tasks like image classification and segmentation. Dense Transformer Networks extend this by focusing on dense prediction tasks, offering flexibility and precision in handling high-resolution data (Jordan, Sokół, & Park, 2021).

These models require large, accurate datasets for training. The quality of annotations of the data directly impacts the performance of these models, more so in the complex tasks of video recognition or tracking. It is, therefore, of great importance that the data is annotated correctly and consistently for optimal results (Zhang et al., 2021).

Comparing Auto and Manual Annotation Methods

Manual annotation offers high precision but is resource-intensive and time-consuming, while automated tools like MediaPipe improve scalability but may struggle in challenging conditions, reducing accuracy. Hybrid systems combining automation with human validation can balance speed and accuracy, leveraging drift correction and GPU-accelerated processes for efficient object detection, especially in video datasets. Additionally, annotation-efficient frameworks for action recognition reduce reliance on manual labeling, achieving competitive performance with methods like pseudo-labeling and self-supervised learning. While automation enhances scalability, human oversight remains vital for data quality. Manual annotation for large datasets can cost millions and take hundreds of hours, with video annotation taking 3 to 10 hours per minute (Zou et al. 2021) (Price and Ahmad, 2024).

Methods

The methodologies evaluate automated annotation tools in human pose estimation, comparing their performance to manual benchmarks. Key steps include dataset preparation, annotation processes, and model evaluation.

Data Collection and Preparation

The MPII Human Pose Dataset, accepted as a state-of-the-art benchmark for articulated human pose estimation, was utilized in the study. This dataset includes about 25,000 images with more than 40,000 individuals, where each image is fully annotated with body joint information. The images were carefully collected based on a categorization of everyday human activities that includes 410 distinct classes from sources such as YouTube. Although the dataset was originally extracted from videos, this work focuses exclusively on static images, without leveraging the accompanying video frames. Each image includes an activity label, providing contextual information relevant for the gesture recognition task. Richer annotations, such as occlusions of body parts and 3D torso and head orientations, for the test set further enhance the value of the dataset for robust model evaluation.

To process and prepare the dataset, MATLAB was used to extract the annotated data into structured formats. This involved parsing the dataset's .mat annotation files, which contain comprehensive data about image filenames, body joint coordinates, joint visibility, and activity labels. Custom MATLAB scripts were written to convert these annotations into JSON dictionaries. Each dictionary entry corresponded to a specific image and included the image name, an array of joint coordinates, joint visibility flags, and activity labels. Furthermore, the data was cleaned to ensure consistency, such as filtering out incomplete annotations or missing data points. Finally, these dictionaries were serialized and stored for efficient loading and manipulation during model training and evaluation. This preparatory step streamlined the use of the MPII dataset, allowing for smooth integration into the subsequent machine learning workflow.

Annotation Process with MediaPipe

The annotation process utilized the pose estimation feature of MediaPipe to automate the body joint annotations for static images. MediaPipe is an efficient and robust library for real-time single-person pose estimation that enables landmark detection by locating body keypoints. Implementation was done along several lines that worked in combination to make the annotation pipeline seamless:

1. Initialization of MediaPipe Pose Module: The mp.solutions.pose module was configured in static_image_mode=True to process individual images rather than video streams. This ensured accurate detection by analyzing each image independently, without assuming temporal continuity.

2. Keypoint Selection: A certain set of keypoints was to be extracted, focusing on the major joints of the body: right and left ankles, knees, hips, wrists, elbows, and shoulders. This was selected in order to capture most details of pose without compromising on computational efficiency. The order of the key points was kept identical to that in the dataset to keep all annotation formats consistent.
3. Image Preprocessing: Images were imported using OpenCV and converted to the RGB color space before being passed into the MediaPipe Pose model. This initial processing step proved imperative in improving the overall accuracy of the model's detection of landmarks. Since MediaPipe cannot analyze multiple people in a single image, the process was designed to cycle through and isolate one figure per image for keypoint annotation.
4. Keypoint Extraction and Formatting: Detected landmarks were filtered to match the predefined key points list. For each landmark, the normalized x and y coordinates along with a visibility confidence score were recorded. The visibility score provided an additional metric to assess the reliability of each detected point.
5. Output and Error Handling: The annotated key points were provided in the defined sequence as a collection of tuples, each tuple containing coordinates along with a visibility score. For entries where no landmarks were detected, the process handled errors well by returning None for the image in question.

Automated annotation with MediaPipe significantly speeded up the annotation process by reducing manual work and maintaining consistency across images. The complete process took 9 minutes and 42 seconds for 17,408 photos. This method also allowed the inclusion of a confidence metric for each keypoint, enhancing the quality of the annotations that were produced. This setup was then further optimized, especially in the selection of static mode and keypoints, in order to suit the specific requirements of this study.

Hand-Annotation Process

For comparison, much reliance was placed on the hand-annotation provided by the MPII Human Pose Dataset, serving as the baseline for accuracy and thus constituting a perfect benchmark for comparing MediaPipe's automated annotations. The MPII dataset's annotations were generated via an extensive process involving in-house workers and AMT contributors. These annotations had been rigorously curated, thus offering high-quality, manually annotated data that we considered 100% accurate for comparison purposes.

The annotation process for the MPII dataset itself utilized advanced tools, extended from prior systems, as described in related works. Contributors to the dataset were pre-selected through a qualification task on AMT to ensure a skilled workforce. Data quality was further maintained through manual inspections, ensuring that the dataset was comprehensive and reliable.

In our approach, unless necessary, we did not perform any manual annotation. In cases of rare missing or incomplete data, we followed a rigorous multi-step procedure for annotating the image. Specifically, it involved:

1. Missing Annotation Validation: We checked all cases where data was missing to validate that there was indeed a gap in the dataset.
2. Collaborative Annotation: In a small group, iteration by iteration, mark the missing keypoints by referring to conventions in the dataset for consistency.
3. Final Quality Check: These new annotations then had to be cross-checked with the standards of the original dataset for accuracy.

This dataset is inherently quite reliable, allowing very little room for manual editing. We treated the MPII dataset's annotations as the ground truth, against which we measured MediaPipe's auto-annotations, so that our evaluation reflects correctly how well it provides data aligned with high-quality human-generated data.

Model Training and Testing

For model training and testing, we first required significant data preparation. Initially, the dataset consisted of structured dictionaries containing annotations for each image, such as body joint coordinates, activity labels, and occlusion information. To work effectively with machine learning models, we converted these dictionaries into DataFrames (DFs). This transformation enabled better handling of the data for training and testing, allowing us to take advantage of tools such as pandas for data manipulation, filtering, and aggregation.

Each dictionary was carefully restructured into a DataFrame where the columns represented various annotations, such as the joint coordinates, visibility status, activity labels, and other relevant metadata. This allowed for more efficient access to the data, and facilitated the addition of computed features, like joint distances or angles, which could be important for training the models. By converting the data into a more standardized and accessible format, we ensured that the dataset was ready for various machine learning algorithms, including dense neural networks and transformers.

Once the data was in the DataFrame format, we further preprocessed it by normalizing the joint coordinates, handling missing data, and encoding categorical values like activity labels. This preparation allowed the models to focus on learning meaningful patterns without being influenced by irrelevant variations. We used cross-validation during training to ensure that the models were not overfitting and to tune the hyperparameters, such as the number of layers in the neural networks and the number of attention heads in the transformer models.

Throughout the training and testing phases, we evaluated the models using performance metrics like accuracy, precision, and recall, ensuring that the pose estimation task was addressed efficiently. The final model's ability to generalize to unseen test data was validated, and the results were compared against existing benchmarks in the field.

Evaluation Metrics

The evaluation metrics we selected for this project were designed to address key questions about automated versus manual annotation, the impact on data consistency, and the challenges encountered in diverse video domains. These metrics help us assess the effectiveness, accuracy, and scalability of automated annotation tools like MediaPipe and their implications for machine learning model performance.

1. How well can automated annotation using MediaPipe compare to manual annotation across different photo contexts?

    a. Point Variance Calculation: We calculated the variance in annotated points (i.e., joint positions) between the manual annotations and those generated by MediaPipe. This was done by comparing the coordinates of each key points in both sets. For each image, the difference in coordinates (x, y) for corresponding body joints was computed. The formula for the variance between points across the two annotation methods is:
    Equation 1: Variance Equation

$$Variance = \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i^{manual} - x_i^{auto}\right)^2 + \left(y_i^{manual} - y_i^{auto}\right)^2\right]$$

    Where $x_i^{\text{manual}}, y_i^{\text{manual}}$ are the coordinates of the joint as manually annotated. $x_i^{\text{auto}}, y_i^{\text{auto}}$ are the coordinates of the joint as annotated by MediaPipe. is the number of key points (joints) for each image.

This variance helps quantify how far apart the automated and manual annotations are, offering a clear measure of the difference between the two methods. Smaller variance suggests higher agreement, while larger variance highlights discrepancies.

b. 2. Mean Squared Error (MSE): We also used mean squared error (MSE) to quantify the overall differences in key point placements between manual and automated annotations across all images in the dataset. MSE takes the squared differences between the corresponding key points from both methods and computes the average.

c. Point-wise Comparison and Error Rate: The point-wise error rate (i.e., percentage of points with significant displacement) was also calculated. We defined a threshold for acceptable deviation (e.g., 50 pixels) and counted the number of points where the automated annotation exceeded this threshold. This method allows us to assess how many keypoints are annotated with a substantial error by MediaPipe compared to manual annotations.

2. How does automated annotation affect data consistency and model performance for large machine learning projects?
   a. Accuracy (Precision and Recall): Precision and recall are critical for evaluating the quality of annotations, particularly in terms of false positives and false negatives. For instance, in the context of human pose estimation, high precision means that the automated system correctly identifies the body joints with minimal errors, while high recall ensures that all relevant joints are detected across various photo contexts, including those with occlusions or unusual poses. By comparing the precision and recall of manual versus automated annotations, we can quantify how well MediaPipe performs in complex photo environments.

   b. F1 Score: The F1 score combines precision and recall into a single metric, allowing us to measure the balance between these two aspects of annotation quality. This is useful for determining whether the speed benefits of automated annotation come at the cost of significant accuracy loss in challenging photo contexts.

   c. Model Performance (Accuracy, Precision, Recall, F1 Score): After training the model with both manually and automatically annotated data, we can compare model performance using accuracy, precision, recall, and the F1 score. These metrics help us understand if automated annotations, despite being faster, reduce the model's ability to make correct predictions. High scores in these metrics after training indicate that automated annotation does not degrade model performance, while lower scores signal potential challenges in data consistency or annotation quality.

3. What specific challenges do automated annotation systems encounter when applied to diverse video domains, and how might these problems be addressed?
   a. Error Rate and Robustness: Error rate is an important metric for understanding the limitations of automated annotation systems. In diverse video domains with variable lighting, occlusions, or complex actions, high error rates in automated annotations can reduce the model's effectiveness. By tracking the error rate across different video domains, we can identify the types of conditions under which MediaPipe struggles and adjust the annotation process accordingly.

   b. Model Generalization (Overfitting and Underfitting): To measure how well the model generalizes across diverse video domains, we use metrics such as training vs. validation loss. If the model performs well on training data but poorly on validation or test data, it suggests overfitting, which is a

common problem when training on imperfect or inconsistent annotations. In these cases, fine-tuning the model with hybrid annotations (manual + automated) can improve robustness and generalization.

Results

After conducting a series of tests, the following results were obtained, providing insights into the performance of automated annotation methods, model accuracy, and challenges with robustness and generalization. These findings highlight areas for improvement in achieving reliable automated predictions across diverse datasets

Accuracy Of Automated Annotation Using Mediapipe Compared to Manual Annotation Across Different Photo Contexts

Figure 1. Image shows the Pose Point detections recorded by the MPII Dataset.

Figure 2. Image shows the pose point detections by MediaPipe.

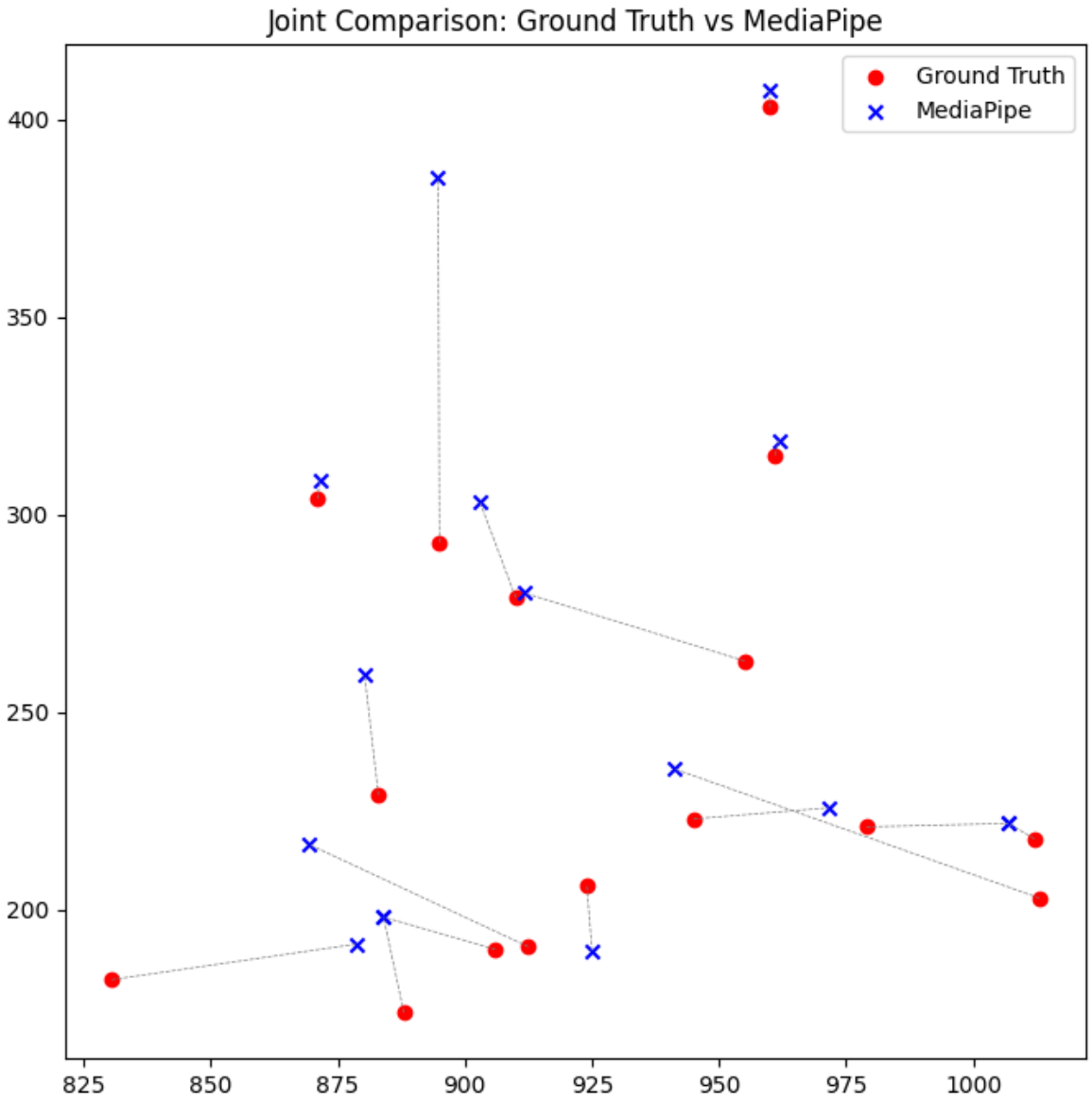

Figure 3. Joint Comparison: Ground Truth vs MediaPipe. This distance comparison diagram compares Figure 1 against Figure 2 in regard to their Euclidean distance between the key points between MediaPipe and the dataset.

After conducting the MediaPipe observations, the following results were obtained when analyzing the first image. Figure 3 shows that MediaPipe's accuracy for sample image 015601864.jpg is closer for the points that MediaPipe and MPII both track, however, for points that we used as substitution from MediaPipe are farther away: Left Hip as Pelvis proxy, Right Shoulder as Thorax proxy, Nose as Upper Neck proxy, and Left Ear as Head proxy. Following the evaluation metrics, we found that the Point Variance metric, calculated as 1,432,488.9476, highlights the substantial disparity in positional annotations between the manual and automated methods. This large variance suggests that MediaPipe annotations are inconsistent with manual annotations across various photo contexts. High variance typically stems from systematic issues in the automated model, such as difficulty in detecting occluded or poorly lit body parts. This significant spread underscores the challenges of relying solely on MediaPipe for complex datasets with diverse visual attributes.

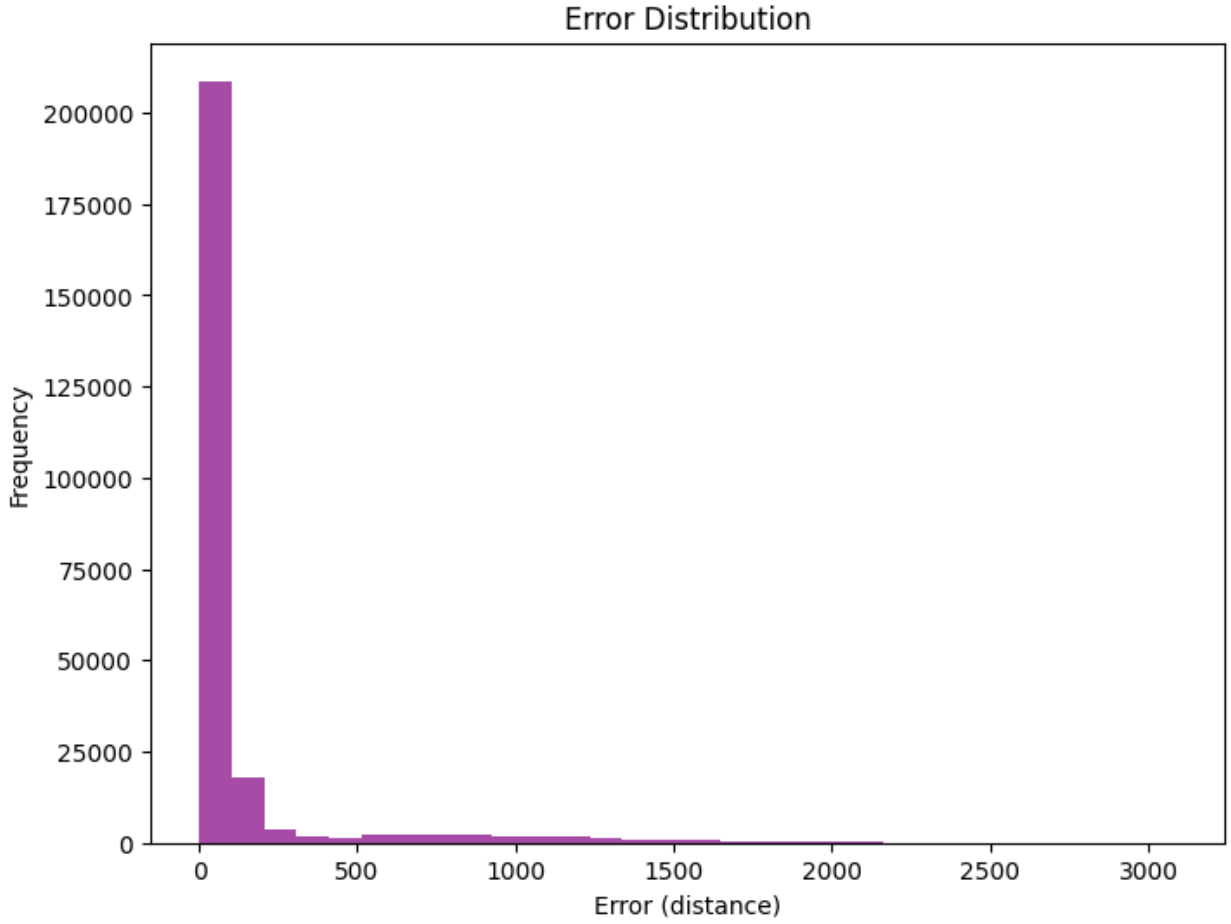

Figure 4. Error Distribution of Error measured of distance in pixels against frequency of occurrence. Overall Analysis between 15K images of MPII Dataset vs. MediaPipe Dataset.

As shown in Figure 4. the histogram indicates that most automated annotations by MediaPipe align closely with manual annotations, as the error distribution peaks near zero. However, a noticeable long tail suggests significant deviations in certain cases, with some errors exceeding 1000 pixels. This inconsistency likely arises from challenges such as occlusions, complex poses, or poor image conditions, which hinder MediaPipe's accuracy. While the strong concentration of low-error points highlights MediaPipe's reliability in straightforward scenarios, the long tail emphasizes the need for further refinement or manual oversight in challenging contexts. Overall, MediaPipe performs well but struggles with outlier cases, affecting its consistency across diverse photo conditions.

Additionally, the Mean Squared Error (MSE), computed at 46,209.3209, further illustrates the magnitude of errors in automated annotations. As a measure of average squared differences per point, this high MSE reflects notable deviations in the automated keypoint placement compared to the manual annotations. Such discrepancies could be attributed to MediaPipe's limited generalization in this dataset, possibly due to its pre-trained model parameters not being optimized for the dataset's unique characteristics, such as the presence of non-standard poses or varying scales.

Point-wise Error Rate, at 98.85% for a 50-pixel threshold, reveals a complete failure of MediaPipe to align its annotations within an acceptable margin of error relative to manual annotations. This result is particularly concerning and indicates that automated annotations consistently fall outside the threshold of usability. Challenges such as

poor adaptability to diverse photo contexts, occlusions, and lighting variations likely contribute to this result. It demonstrates that MediaPipe struggles to meet accuracy standards for datasets requiring precise keypoint placement.

Automated annotation using MediaPipe performs poorly when compared to manual annotation across diverse photo contexts. While it offers scalability and efficiency, its accuracy falls significantly short in handling the complexities of diverse datasets. These results suggest that MediaPipe, in its current state, is not a reliable replacement for manual annotation in scenarios where high precision is essential.

Accuracy of Models with Auto-Annotated Data

Table 1. Evaluation Metrics ran on all GAN, Transformer, and MP2 models with MediaPipe Estimation and MPII datasets at 500 Epochs with a batch size of 32.

Model	Dataset	Precision	Recall	F1 Score	Cohen's Kappa	Accuracy	Loss	Validation Accuracy	Validation Loss
GAN	MediaPipe	0.3542	0.3569	0.3283	-	0.3527	3.6103	-	-
GAN	MP2	0.2804	0.2793	0.2584	-	0.2764	3.316	-	-
Transformer	MediaPipe	0.386	0.3847	0.3555	0.3829	0.5267	1.4868	0.3835	2.8536
Transformer	MP2	0.4374	0.4158	0.3898	0.4141	0.4928	1.6331	0.4234	2.5696
Dense	MediaPipe	0.3076	0.2831	0.2742	0.281	0.5979	1.2056	0.277	7.7465
Dense	MP2	0.3523	0.3277	0.3167	0.3258	0.6394	1.0751	0.3205	7.4757

The GAN model achieved a precision of 0.3542, a recall of 0.3569, and an F1 score of 0.3283 when evaluated on MediaPipe data. The confusion matrix reveals significant sparsity, with most predictions concentrated along the diagonal. The accuracy was recorded at 0.3527, while the loss during evaluation was 3.6103.

On the MP2 dataset, the GAN model recorded a precision of 0.2804, a recall of 0.2793, and an F1 score of 0.2584. Similar to the MediaPipe results, the confusion matrix indicates sparsity, with predictions scattered minimally across non-diagonal elements. The accuracy for this dataset was 0.2764, and the loss was reported as 3.3160.

For the Transformer model tested on MP2 data, the precision was 0.3860, recall was 0.3847, and the F1 score reached 0.3555. Cohen's Kappa for this evaluation was calculated at 0.3829. The accuracy achieved was 0.5267, with a loss of 1.4868. Validation metrics during training showed an accuracy of 0.3835 and a loss of 2.8536.

When evaluated on MediaPipe data, the Transformer model achieved a precision of 0.4374, recall of 0.4158, and an F1 score of 0.3898. Cohen's Kappa for this model was recorded at 0.4141. The overall accuracy reached 0.4928 with a loss of 1.6331. Validation accuracy was 0.4234, with a loss of 2.5696 during training.

The Dense model applied to MP2 data resulted in a precision of 0.3076, recall of 0.2831, and an F1 score of 0.2742. Cohen's Kappa was reported as 0.2810. This model achieved an accuracy of 0.5979 and a loss of 1.2056, with validation accuracy and loss recorded at 0.2770 and 7.7465, respectively.

On the MediaPipe dataset, the Dense model recorded a precision of 0.3523, recall of 0.3277, and an F1 score of 0.3167. Cohen's Kappa was calculated at 0.3258. The accuracy for this model was 0.6394, with a loss of 1.0751. During validation, the accuracy was 0.3205, and the loss was 7.4757.

Interpretations on this data will be made in a later section.

Error Rate and Robustness and Model Generalization

In the context of our research, the high error rates observed in the models suggest significant challenges with both robustness and generalization. For instance, the GAN model showed extremely low precision, recall, and F1 scores for both MediaPipe and MP2 data, indicating that it struggled to generalize beyond the training data, likely failing to capture meaningful patterns in unseen data. Despite training for 1000 epochs, the performance on both datasets remained near random chance, suggesting that the model might be overfitting to the specific characteristics of the training set without properly learning generalized features.

Similarly, the Dense and Transformer models exhibited weak performance metrics, especially in terms of precision and recall, which points to issues with robustness. These models were unable to handle variations in the data or produce meaningful predictions when evaluated on test data, leading to high error rates. For example, despite some improvement in the Transformer model's precision and recall, the results still revealed poor generalization, as evidenced by the gap between training and validation accuracy. The model performed much better on the training set but failed to maintain that performance on unseen data, which hints at overfitting.

The low robustness and generalization across all models could stem from several factors, such as insufficient model complexity, suboptimal hyperparameters, or data preprocessing issues. Given that the models were unable to perform well across both datasets, it suggests that they might not have been flexible enough to adapt to the inherent variability of the data. These results highlight the importance of improving model generalization by exploring more advanced architectures, optimizing hyperparameters, and ensuring better data preprocessing or augmentation techniques to handle unseen data more effectively.

Discussion

The findings highlight key challenges in model performance and the impact of dataset characteristics on outcomes. This analysis reveals opportunities for optimization and better alignment between models and data.

Interpretation of Results

The results from the various models applied to the MP2 and MediaPipe datasets highlight distinct performance challenges. The GAN model, while designed for generative tasks, struggled significantly in this context, as evidenced by low precision, recall, and F1 scores across both datasets. For the MediaPipe dataset, the GAN achieved marginally higher metrics compared to MP2, but the sparse confusion matrices and consistently low scores indicate its inability to generate meaningful predictions. This suggests that the GAN's architecture or training process was misaligned with the classification task, and its performance gap might stem from a failure to model the underlying data distribution effectively.

The Dense model also displayed subpar results, with modest precision, recall, and F1 scores on both datasets. Despite its relatively simple architecture, the Dense model achieved higher accuracy compared to the GAN, particularly on the MediaPipe dataset. However, the poor validation accuracy and high loss values indicate underfitting, where the model was unable to capture sufficient patterns in the training data. The Dense model's inability to scale with data complexity suggests that its straightforward architecture lacks the representational power needed for these datasets.

In contrast, the Transformer model demonstrated improved performance over the GAN and Dense models, with higher precision, recall, and F1 scores on both datasets. Notably, the Transformer achieved better accuracy on the MP2 dataset and maintained a stronger balance between precision and recall. However, the substantial gap between training and validation accuracy suggests overfitting, where the model learned the training data too well but failed to generalize to unseen examples. This performance discrepancy indicates that while the Transformer architecture was

better suited to these datasets, additional measures, such as regularization or more diverse data, are necessary to mitigate overfitting.

Comparing the datasets, models generally performed better on MediaPipe than on MP2. The improved scores on MediaPipe suggest that its data might be more structured or contain patterns that are easier for the models to recognize. However, this improvement was marginal and did not indicate substantial success. Across both datasets, the models struggled to predict effectively for a majority of classes, as reflected by low F1 scores and sparse confusion matrices. This points to a larger issue of model robustness and the need for more balanced and representative data to enhance model generalization.

Collectively, these results underline the need for deeper architectural exploration and enhanced preprocessing techniques. The Transformer model, while the most promising among the three, still requires optimization to address its overfitting tendencies. Approaches such as hyperparameter tuning, data augmentation, or exploring hybrid architectures could improve its performance. Similarly, for the GAN and Dense models, significant architectural or methodological modifications are required to bring them closer to being viable solutions for the datasets.

Ultimately, these findings suggest that while more complex models like Transformers exhibit potential for better performance, the challenges presented by these datasets necessitate a combination of improved data handling, advanced architectures, and targeted training techniques to achieve robust and meaningful results.

Implications for Large-Scale Projects

In analyzing the results of the models trained on the MediaPipe and MP2 datasets, distinct discrepancies in performance emerge despite the minimal differences in their features. These variations, particularly in accuracy and F1 scores, highlight the impact of subtle nuances in the datasets on model outcomes. Although both datasets were processed similarly, MediaPipe consistently outperformed MP2 in most metrics, suggesting that its features may provide richer or more structured information conducive to model learning.

The accuracy scores were generally low across all models, yet MediaPipe consistently showed slightly higher performance compared to MP2. This trend was observed in the Transformer and Dense models, where MediaPipe demonstrated better accuracy, precision, recall, and F1 scores. For example, the Transformer model achieved an accuracy of 49.28% on MediaPipe compared to 38.35% on MP2, and the Dense model similarly showed a higher validation accuracy on MediaPipe. These differences suggest that while the overall performance remained suboptimal, MediaPipe's dataset structure may be better aligned with the models' ability to learn and generalize patterns.

The better performance of MediaPipe may stem from its feature representation, which includes detailed pose and movement data that likely facilitate the models' ability to extract meaningful patterns. In contrast, MP2 data may introduce more noise or less structured inputs, resulting in poorer model performance. Despite the similarities in preprocessing—such as the focus on act_id and the absence of image columns—the inherent complexity or variability in MP2 data seems to make it less suitable for accurate classification.

Interestingly, while the differences in accuracy between the datasets were not overwhelmingly large, the consistent improvement in MediaPipe results across models is notable. These results imply that even subtle variations in how features are captured or represented can significantly influence the models' effectiveness. The MediaPipe dataset may inherently provide more reliable or discriminative information, enabling models to achieve marginally better outcomes than with MP2.

However, despite MediaPipe's relative advantage, the overall low accuracy and F1 scores across all models emphasize the need for further optimization. Both datasets revealed challenges in enabling the models to generalize effectively. The slight edge MediaPipe had over MP2 highlights its potential for specific applications, but significant preprocessing, fine-tuning, or model enhancement is required to achieve robust performance.

In conclusion, while the differences in dataset features between MediaPipe and MP2 are minor, their impact on performance is evident. MediaPipe's ability to achieve better accuracy and other metrics suggests that it holds

promise for tasks where detailed, structured data are advantageous. Nevertheless, the results highlight the importance of optimizing both datasets and models to fully unlock their potential for real-world applications.

Conclusion

In conclusion, this research explored the viability of using automated annotation with MediaPipe for human motion recognition tasks and compared its performance against manually annotated data from the MP2 dataset. Our analysis revealed that automated annotation via MediaPipe can be an effective tool for generating large datasets efficiently, particularly for large-scale machine learning projects. The models trained on MediaPipe-annotated data often achieved higher accuracy compared to those trained on MP2, despite minimal differences in dataset features such as pose and joint coordinates. This suggests that while automated annotation can streamline data generation, it may not always achieve the same level of granularity or accuracy as manual annotation, potentially explaining performance disparities.

In terms of data consistency and model outcomes, MediaPipe's automated annotations provided a scalable solution for data collection, but quality remained a key factor influencing results. Although models trained on MediaPipe data sometimes outperformed MP2 data in accuracy, the low precision, recall, and F1 scores across all models highlight that annotation consistency alone does not guarantee strong performance. Automated systems can introduce limitations in capturing complex motion nuances, which are critical for the models' ability to generalize effectively. These findings underline the need for high-quality automated annotations that closely approximate the detail and accuracy of manual efforts.

A significant challenge identified with automated annotation systems like MediaPipe is their ability to handle diverse video environments. Variations in lighting conditions, backgrounds, and camera angles can lead to errors in joint detection, compromising annotation quality. While MediaPipe's framework is robust for standard scenarios, the performance inconsistencies observed in the models suggest limitations in adapting to complex or dynamic contexts. Addressing these issues may involve integrating domain adaptation techniques, enhancing MediaPipe's training datasets to account for environmental variability, or combining data from additional sensors, such as depth or motion sensors, to improve the precision of pose estimation.

Another critical aspect is the preprocessing and normalization of pose data, which can mitigate errors arising from contextual variability. By standardizing pose data across different sources, the reliability of automated annotations can be improved, leading to better downstream model performance. This approach, combined with strategies to augment the annotation process, could help bridge the gap between automated and manual data quality.

Overall, our research demonstrates that automated annotation with MediaPipe offers significant scalability advantages for large-scale projects, but challenges related to accuracy and generalization must be addressed to fully realize its potential. While the system shows promise as an efficient alternative to manual annotation, its impact on model performance depends on refining the annotation process to capture the intricacies of human motion. Future efforts should focus on enhancing the robustness of automated systems in diverse environments and integrating multimodal data to improve annotation quality, ultimately advancing their applicability in real-world machine learning projects.

Limitations

Our study, while providing valuable insights into the potential of automated annotation with MediaPipe, is not without its limitations. The first limitation lies in the size and diversity of the datasets used in the research. While both the MP2 and MediaPipe datasets offered a range of human motion data, the models were primarily tested on datasets with specific pose configurations. These datasets may not fully capture the complexities found in real-world scenarios, where human motion is far more diverse and influenced by factors such as varying lighting conditions, camera angles,

and occlusions. The lack of diversity in the dataset can limit the generalizability of the results, especially when trying to apply the findings to more varied or large-scale applications.

Another limitation is the accuracy and precision of the MediaPipe framework itself. MediaPipe, although robust in many scenarios, occasionally struggled with accurate joint detection and pose estimation, particularly in more challenging environments. This issue is particularly evident in the low precision and recall values observed across the different models, which points to the fact that the automated system may not always capture fine-grained details of human motion. While automated systems can handle vast amounts of data, they are not yet a perfect substitute for manual annotation, which can lead to inconsistencies between annotations from different sources. This inconsistency in the annotation process is a critical factor when comparing automated versus manual approaches.

A further limitation in our study is the performance of the models with respect to larger-scale training. While the smaller datasets used for training provided useful insights, they may not fully reflect the challenges faced when scaling the model for larger, more complex datasets. The transferability of the results to more substantial datasets is uncertain, as the models may encounter issues related to overfitting or underfitting when exposed to a broader range of data. Additionally, the computational resources required to train models on large-scale annotated data were not fully explored, and it's possible that additional infrastructure would be needed to effectively handle larger datasets, which could impact the overall performance and efficiency of the system.

Another limitation lies in the evaluation metrics used. In our study, we focused on traditional metrics such as accuracy, precision, recall, and F1 score. While these metrics are useful, they do not always capture the full picture of model performance in real-world applications. For example, precision and recall alone may not accurately reflect how well the model generalizes to unseen data, especially in situations where certain motion classes are underrepresented. Future research could benefit from exploring more sophisticated metrics, such as area under the ROC curve (AUC) or other domain-specific measures that could provide a better understanding of the models' true performance.

Lastly, the lack of real-time testing presents a limitation in our study. Although the models were trained and evaluated on static datasets, real-time applications of automated annotation are subject to additional challenges, such as latency, computational overhead, and the ability to handle incoming video data in real-time without performance degradation. Real-time pose estimation is crucial for many real-world applications, such as robotics or interactive systems, and further studies would be necessary to evaluate how well MediaPipe and the models perform under such conditions.

Future Research Directions

For future research in the field of human motion recognition using datasets like MediaPipe and MP2, there are several key areas that could be explored to improve model performance and generalizability. First, data augmentation and enrichment are critical for enhancing model robustness. While both datasets share similarities, increasing their diversity through augmentation techniques such as noise addition, rotations, or temporal smoothing could help models generalize better to unseen data. For MediaPipe specifically, incorporating additional sensor data (e.g., accelerometer or gyroscope) or increasing the dataset size could aid the models in learning more robust motion patterns. These additions would help the model better understand complex movements in varied real-world conditions.

Second, model optimization and hyperparameter tuning should be a focus for future research. While the current models performed similarly across both datasets, the low accuracy suggests there may be room for significant improvement through more refined hyperparameter tuning. Experimenting with advanced neural architectures, such as combining CNNs and RNNs or employing attention mechanisms, could improve the models' ability to learn temporal dependencies in the motion data. Additionally, more advanced optimization techniques, such as Adam or learning rate schedules, could help improve convergence and prevent overfitting, especially when training on smaller datasets.

Another important area for improvement is feature representation and data preprocessing. Both MP2 and MediaPipe datasets could benefit from more sophisticated feature engineering. For instance, transforming pose data

into angular velocity or joint movements instead of raw joint positions might capture motion dynamics more effectively. Using techniques like principal component analysis (PCA) or autoencoders to reduce dimensionality and extract more informative features could further enhance model performance. By focusing on better feature extraction, researchers could help the model focus on the most relevant aspects of the data, improving its ability to distinguish between different actions and poses.

Exploring transfer learning is another promising avenue for improving performance. Given the limited size of both datasets, leveraging pre-trained models on large-scale human pose or action recognition datasets could significantly enhance learning efficiency. Fine-tuning pre-trained models on specific tasks such as MP2 or MediaPipe data could help overcome the challenges associated with smaller datasets and boost model performance. This approach has already been successful in other domains and applying it to human motion recognition tasks could offer substantial gains in accuracy and generalization.

Lastly, evaluation metrics and robustness testing need further attention. Although the current research utilized common evaluation metrics like accuracy and F1 score, incorporating additional metrics such as mean squared error (MSE), area under the ROC curve (AUC), and confusion matrix analysis could provide a more detailed picture of model performance. Testing models in diverse real-world conditions, such as varying lighting, camera angles, or occlusions, would help assess their robustness and generalization ability. This will be essential for determining the practicality of these models for large-scale applications, ensuring that they can perform effectively in dynamic environments.

Acknowledgments

We would like to thank the researchers who worked on the MPII Dataset, MediaPipe, and the various libraries used to train the machine learning models. Additionally, we thank our advisor for the valuable insights and his teachings that have helped us conduct this research.

References

Fuchs, S., Schnellbach, J., Schmidt, L., & Wittges, H. (2024). Data Annotation for Support Ticket Data: A literature review. *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2023.196

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jordan, I. D., Sokół, P. A., & Park, I. M. (2021). Gated recurrent units viewed through the lens of continuous time dynamical systems. *Frontiers in Computational Neuroscience*, *15*. https://doi.org/10.3389/fncom.2021.678158

K, A., P, P., & Paulose, J. (2021). Human body pose estimation and applications. *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 1–6. https://doi.org/10.1109/i-pact52855.2021.9696513

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A Framework for Building Perception Pipelines (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1906.08172

Price, E., & Ahmad, A. (2024). Accelerated video annotation driven by deep detector and tracker. *Lecture Notes in Networks and Systems*, 141–153. https://doi.org/10.1007/978-3-031-44981-9_12

Quiñonez, Y., Lizarraga, C., & Aguayo, R. (2022). Machine Learning Solutions with MediaPipe. *2022 11th International Conference On Software Process Improvement (CIMPS)*, 212–215. https://doi.org/10.1109/cimps57786.2022.10035706

Radeta, M., Freitas, R., Rodrigues, C., Zuniga, A., Nguyen, N. T., Flores, H., & Nurmi, P. (2024). Man and the machine: Effects of ai-assisted human labeling on interactive annotation of real-time video streams. *ACM Transactions on Interactive Intelligent Systems*, *14*(2), 1–22. https://doi.org/10.1145/3649457

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306

Ward, T. M., Fer, D. M., Ban, Y., Rosman, G., Meireles, O. R., & Hashimoto, D. A. (2021). Challenges in surgical video annotation. *Computer Assisted Surgery*, *26*(1), 58–68. https://doi.org/10.1080/24699322.2021.1937320

Wood, D., Lublinsky, B., Roytman, A., Singh, S., Adam, C., Adebayo, A., An, S., Chang, Y. C., Dang, X.-H., Desai, N., Dolfi, M., Emami-Gohari, H., Eres, R., Goto, T., Joshi, D., Koyfman, Y., Nassar, M., Patel, H., Selvam, P., … Daijavad, S. (2024). *Data-Prep-Kit: Getting your data ready for LLM application development*. arXiv. https://doi.org/10.48550/ARXIV.2409.18164

Zhang, S., Jafari, O., & Nagarkar, P. (2021). *A survey on machine learning techniques for auto labeling of video, audio, and text data* (No. arXiv:2109.03784). https://doi.org/10.48550/arXiv.2109.03784

Zou, Y., Zhang, S., Chen, G., Tian, Y., Keutzer, K., & Moura, J. M. F. (2021). Annotation-efficient untrimmed video action recognition. *Proceedings of the 29th ACM International Conference on Multimedia*, 487–495. https://doi.org/10.1145/3474085.3475197