ChromaPose: Robustness of 2D Pose Estimation Under Different Color Illuminations

Jamiu Oluwaseun Ojeleye[1], Pratik Singh Bisht[1], Philippe Colantoni[2], Damien Muselet[2], and Alain Tremeau[2]

Laboratoire Hubert Curien, UMR 5516
(jamiu.ojeleye,pratik.bisht)@univ-st-etienne.fr
(philippe.colantoni,damien.muselet,alain.tremeau)@univ-st-etienne.fr

Abstract. This study investigates the robustness and precision of 2D human pose estimation techniques, particularly feature-based and 3D shape-based methods, against the backdrop of varying color illumination conditions simulated by chromatic adaptation, as well as varying spatio-temporal dynamics. While the AIST++ dance videos dataset serves as the primary data, the insights gained are pertinent to broader contexts like sports, highlighting the critical influence of lighting on the performance of pose estimation technologies.

Keywords: 2D Pose estimation · Chromatic Adaptation · Lighting Conditions · Feature-Based Pose Estimation · Shape-Based Pose Estimation.

1 Introduction

ChromaPose aims to provide tools and methods for the analysis of dance and theatre performances, particularly from archival video materials. ChromaPose seeks to identify and address common challenges in archival material, such as poor color and light contrast and inadequate lighting, which can significantly impede the visibility of performers' movements and consequently, the precision of human body pose estimation. At the heart of ChromaPose's research is the AIST++ dataset [23], a rich collection of dance videos showcasing a variety of styles and viewpoints. Despite its breadth, the dataset does not fully represent the challenges typical in the digitization of performing arts, such as the effects of stage lighting, shadows, and costume colors on visual clarity.

To address these gaps, ChromaPose subjects the AIST++ dataset to specific illumination changes simulated by chromatic adaptations to explore how changes in color and brightness affect the performance of 2D pose estimation techniques. The project aims to identify the most effective 2D pose estimation methods that can withstand the dynamic lighting and chromatic shifts often found in archival dance and theatre footage, as illustrated in Fig. 1, and by developing a novel visualization technique to evaluate the accuracy and stability of these methods. This innovative visualization not only enables a comparative analysis of different techniques but also enhances our comprehension of their performance amidst the unique challenges of the performing arts. Furthermore, by delving into the complexities of temporal performance footage, the research expands the scope of digital archiving and human pose estimation, highlighting solutions to visual challenges and setting the stage for creating new strategies that take into account lighting and visual factors for future data collection initiatives.

Fig. 1. Sample challenging scenarios: strong colored illumination, complex poses, occlusion, low light and back lit conditions, in archive videos from PREMIERE project [20].

2 Related Works

The AIST++ Dance Motion Dataset [23], an extension of the AIST Dance Video Database, significantly advances dance movement analysis and human pose estimation by utilizing multi-view videos to estimate camera parameters, 3D human keypoints, and dance motion sequences. With over 10.1 million images of 30 subjects from 9 views, providing the largest collection of 3D human keypoint annotations, and encompassing 1,408 sequences across 10 dance genres, AIST++ supports diverse research tasks like multi-view keypoints estimation, motion prediction/generation, and cross-modal analysis with music. This dataset, featuring professional dancers in various styles, enhances human pose estimation models, showing marked improvements in accuracy and robustness, especially against complex movements and occlusions. Its real-world scenarios and detailed annotations make AIST++ invaluable for developing new methodologies in pose estimation and bridging computer vision with performing arts, thus serving as a key resource for advancing research in these fields.

2.1 Human Detection and Tracking

In the realm of human body detection and tracking, the integration of state-of-the-art detection algorithms with advanced tracking methodologies forms the backbone of current systems. YOLOv7 [24] exemplifies this integration, offering real-time, high-accuracy object detection. This version enhances the capabilities of its predecessors through architectural optimizations and algorithmic improvements, making it exceptionally suitable for dynamic and complex environments where rapid object detection is crucial. Paired with YOLOv7, ByteTrack [28] emerges as a robust tracking system that exploits the detection efficiency of YOLOv7 to offer seamless tracking across video frames. ByteTrack effectively addresses the challenges of occlusions and rapid object movements by employing a sophisticated association strategy. This approach leverages spatial-temporal information to maintain consistent object identities, thereby significantly reducing identity switches and enhancing tracking continuity. StrongSORT++ [5] introduces filtering and Gaussian Smooth Interpolation (GSI) techniques to further refine tracking accuracy. The filtering relies on probabilistic models to predict future object states based on historical movement data, which is instrumental in handling the unpredictability of object trajectories. This predictive capability is crucial for maintaining the track of objects when they move swiftly or become momentarily obscured. GSI smooths the estimation of object trajectories. By assuming object movements follow a Gaussian distribution, it interpolates positions during periods of occlusion or lost detections, ensuring a coherent and accurate representation

of object paths over time. This method not only aids in overcoming interruptions in object visibility but also contributes to a more stable tracking.

2.2 Feature based 2D Pose Estimation

Feature-based 2D pose estimation is a pivotal process which aims to detect and track human body parts within images and videos, fostering advancements in interactive technologies and motion analysis. Human pose estimation methodologies bifurcate into two main approaches: bottom-up and top-down, each adopting a unique perspective in identifying and associating human body parts. Bottom-up Approaches approaches address pose estimation as a part of detection challenge. OpenPose [2] leverages a complex Convolutional Neural Network (CNN) architecture to detect individual body parts initially. It introduces Part Affinity Fields (PAF), innovatively predicting the connections between body parts to elucidate the human body's structure in imagery. Contrarily, top-down methods such as Convolutional Pose Machines (CPMs) [25] and AlphaPose [7] (or RMPE for Regional Multi-person Pose Estimation) begin with an aggregate human body model, refining it to align with observed image configurations. AlphaPose utilizes pre-established human body models to identify essential keypoints, laying the groundwork for pose reconstruction. While another approach implied by MIPNet [11] is notably proficient in managing occlusions and variations in human body scale, posture, and appearance, ensuring resilient pose estimation.

Amidst these established methods, ViTPose [26] [27] emerges at top, distinguishing itself by integrating Vision Transformer (ViT) technology [4] with traditional pose estimation frameworks. Introduced in 2022, ViTPose revolutionizes the analysis of visual data by adopting transformers to meticulously model the spatial relationships among body parts. This approach excels in managing long-range dependencies, thereby enhancing the accuracy and efficiency of pose estimation. Demonstrating superior performance over predecessors like OpenPose [2] and AlphaPose, as demonstrated in Table 1 (a), ViTPose sets new benchmarks in the field. Its architecture, optimized for deep and nuanced feature extraction, outshines conventional CNN-based methods, proving to be more adaptable and precise across a variety of visual contexts.

2.3 Shape based 2D Pose Estimation

The advent of shape-based 2D pose estimation represents a breakthrough in computer vision, using 3D models to analyze and project human forms onto 2D planes. This approach integrates human anatomy with mathematical models to reconstruct poses accurately, addressing challenges like occlusions and body size variations. Central to this method is the Skinned Multi-Person Linear (SMPL) [16] model, which simplifies capturing the human body's geometry from 2D images, enabling advances in pose estimation. 4DHumans [8] introduces a revolutionary approach to human mesh recovery and tracking over time. At its core is HMR 2.0 [10], a transformer-based network that sets new standards in analyzing and reconstructing human poses from single images, even those previously deemed challenging. When applied to video, it employs 3D reconstructions from HMR 2.0 [10] for advanced tracking capabilities, managing multiple

individuals and maintaining their identities through occlusions. 4DHumans [8] exemplifies state-of-the-art tracking performance in monocular video analysis, also showcasing significant improvements in action recognition tasks. OSX [14] shifts the focus towards a holistic recovery of the 3D human body, including face and hands, from single images. By overcoming resolution challenges without resorting to segmented network processing for each body part, OSX presents a unified pipeline named Component Aware Transformer (CAT). This model employs a body encoder for overall parameter prediction and a local decoder for high-resolution face and hand details, ensuring cohesive and realistic human posture and movement representations. The introduction of the UBody dataset further underscores OSX's applicability to real-world scenarios, highlighting its adaptability and precision in expressive whole-body mesh recovery. SMPLer-X [1] ventures into Expressive Human Pose and Shape estimation (EHPS) by proposing a generalist foundation model that employs a vision transformer (ViT) architecture up to the ViT-Huge variant. It capitalizes on a massive training corpus from a wide array of datasets, demonstrating remarkable versatility and accuracy across numerous testing environments. SMPLer-X's [1] methodology includes rigorous data and model scaling, highlighting the critical role of dataset diversity and model capacity in enhancing EHPS performance. The model's success is further evidenced by its state-of-the-art results on several benchmarks, including AGORA [19] and UBody [14], underscoring the potential of vision transformers in complex human body pose estimation tasks.

These models underscore the transformative potential of shape-based 2D pose estimation. By bridging the gap between 2D imaging and 3D structural understanding, they pave the way for innovations across various domains. From animation and virtual reality to healthcare and sports analytics, the implications of accurately capturing the human body form in motion are profound and far-reaching.

3 Methodology

3.1 Human Detector and Pose Filtering

Within the scope of the PREMIERE project [20], we have crafted a pipeline that addresses various digital processing stages, with a key feature being the initial separation of the 'human detector' from subsequent stages like human body pose estimation and segmentation. This separation aims to enhance detection accuracy and ensure spatial-temporal consistency vital for video analysis, with additional algorithms discussed later. This approach also isolates the accuracy and error rates of human body detection from pose estimation, allowing for an impartial comparison of different pose estimation techniques on our chromatically adapted dataset. We employed YOLOv7 [24] E6E variant, selected for its support of higher resolution (1280x1280 pixels) and reliability in person detection. However, to mitigate false positives in challenging conditions such as low resolution, rapid movement, and insufficient lighting, which could erroneously classify humans as other entities (e.g., dogs, cats, chairs), we restricted the detection to the 'person' class exclusively.

To complement the human body detection framework, the integration of a robust tracking algorithm was deemed necessary. ByteTrack [28] was selected for its efficacy and compatibility with YOLOv7 [24], demonstrating reliable tracking performance

(a)

Method	COCO		CrowdPose		OCHuman	
	val	test	val	test	val	test
Top-Down Methods, ResNet101 + YOLO-v3						
MaskRCNN [9]	-	64.8	-	57.2	-	20.2
AlphaPose [7]	-	70.1	-	61.0	-	-
CrowdPose [13]	-	70.9	-	66.0	-	-
OPEC-Net [21]	-	73.9	-	70.6	-	29.1
MIPNet [11]	72.2	74.2	63.4	68.1	32.8	35.0
Top-Down Methods, HRNet-W48 (384x288) + Faster R-CNN						
HRNet [22]	76.3	75.5	68.0	69.3	37.8	37.2
MIPNet [11]	76.3	75.7	68.8	70.0	42.0	42.5
Top-Down Methods, ViT based						
VitPose++-H [26] [27]	**79.4**	-	-	-	**85.7**	-
Bottom-Up Methods						
OpenPose [2]	-	64.2	-	-	-	-
HigherHRNet [3]	67.1	70.5	-	67.6	-	-

(b)

Setting	**Target(r, g, b)**	**Luminance**
1	$(1.0, 0.3, 0.3)$	-10
2	$(0.63137, 1.0, 0.4823529)$	-10
3	$(0.2, 0.3, 1.0)$	-10
4	$(1.0, 0.3, 0.3)$	0
5	$(0.63137, 1.0, 0.4823529)$	0
6	$(0.2, 0.3, 1.0)$	0
7	$(1.0, 1.0, 1.0)$	-15

Table 1. (a) Comparison of top-down and bottom-up state-of-the-art methods [18]; (b) Color illumination conditions settings.

across standard datasets like MOT17 [17]. However, its application to archival footage revealed limitations, in challenging scenarios like low-light, low resolution, occlusion with very low features, prompting the exploration of advanced tracking capabilities. This exploration led to the incorporation of algorithms from StrongSORT++ [5], which, while independently capable, showed that the default ByteTrack [28] tracking outperformed the base version of StrongSORT in the context of archival footage. Consequently, with some reference to previous works [6] in comparisons and combinations of these methods, a hybrid approach called ByteTrack++ was developed, combining ByteTrack's [28] default tracking capabilities with enhancements from StrongSORT++ algorithms, namely AFLink and GSI, to achieve superior tracking performance.

Furthermore, to achieve more accurate and stable tracking of human poses over temporal video sequences, we implemented and applied a pose filtering method to each estimated pose keypoints. This pose filtering method involves applying Kalman filter to remove high noise levels from the estimated poses generated by the pose estimators followed by applying the Gaussian-based interpolation method proposed in [28] for continuous and visually coherent trajectories of human poses over time.

3.2 Change of color illumination conditions

Chromatic adaptation (CA) can be used for simulating strong changes of color appearance to suit different color lighting conditions. In this research, we employed the Bradford chromatic adaptation method [12] due to its ability to mimic the adaptive mechanisms of the human visual system. Furthermore, we investigated the impact of employing four different destination white points alongside three distinct luminance settings (refer Table 1 (b)) using this method to augment AIST++ [23] video frames with various color illumination conditions.

Estimation of the Illuminant: The estimation of the illuminant for each frame relies on the principle that in scenes lacking dominant light sources, the average color tends towards gray. This estimation provides a crucial reference point for subsequent CA processes.

Fig. 2. Example of simulated color illumination shifts. From left to right, top to bottom: Original AIST frame [23], resulting color shifts for settings 1 to 7.

Chromatic Adaptation with Varying Destination White Points: To explore the effect of different destination white points on human body pose estimation, we simulated color illumination shits using various target white points. This process involved several key steps: (a) Conversion of the source and destination white points from sRGB to XYZ color space, a tristimulus model representing human color perception. (XYZ values were normalized to ensure consistency in luminance) ; (b) Utilization of the Bradford CA matrix to transform XYZ values to LMS (Long, Medium, Short) color space ; (c) Calculation of the LMS gain by scaling destination LMS with source LMS, then multiplying CA matrix to obtain the CA transform. ; (d) the CA transform was then applied on the source image in the XYZ color space.

Luminance Adjustment: Further processing involved the transformation of the resulting XYZ color space into CIELAB color space, effectively separating chromatic information (a*, b*) from luminance (L*). The luminance values in the CIELAB color space were subsequently adjusted to emulate variations in lighting conditions, enabling the simulation of different illuminants' effects on image appearance. Finally, following luminance adjustment, the CIELAB values were converted back into sRGB, completing the color illumination shift process and yielding simulated images adapted to diverse color and luminance conditions. See examples of simulated color shifts in Fig. 2.

3.3 Metrics

Object Keypoint Similarity (OKS): As proposed in the MS-COCO dataset [15], the Object Keypoint Similarity (OKS) measure serves as a fundamental metric for assessing the accuracy of 2D human pose estimation. It quantifies the similarity between predicted and ground truth keypoint locations. In our research, we employ OKS to evaluate the accuracy of various estimation methods. Specifically, we evaluate accuracy using both the average OKS across all joints and the per joint OKS. This dual assessment approach offers several advantages. By considering the average OKS across all joints, we obtain a holistic view and insights into the overall accuracy and robustness of the various pose estimation methods' across the entire pose. We can identify which methods are more resilient to changes in color illumination and which may exhibit inconsistencies.

In addition to the aggregate evaluation, we also assess performance at the individual joint level using per joint OKS. This allows us to pinpoint specific joints where a pose estimation method excels or struggles. Mean Absolute Error (MAE) is another commonly used metric to quantify the accuracy of joint location estimation in 2D human pose estimation. In our research, we utilize MAE to evaluate the accuracy and consistency of various pose estimation method in estimating both the x-coordinate MAE_x and y-coordinate MAE_y for each individual joint.

$$MAE_{xj} = \frac{1}{N}\sum_{i=1}^{N}|x_{ij} - x_{ij}^{GT}| \quad MAE_{yj} = \frac{1}{N}\sum_{i=1}^{N}|y_{ij} - y_{ij}^{GT}|$$

The mean absolute error for the x-coordinate of the joint (resp. for the y-coordinate) is denoted by MAE_x (resp.MAE_y). Here, N is the total number of samples, x_{ij} and y_{ij} are the predicted x and y coordinates for joint j in the ith sample respectively, and x_{ij}^{GT} and y_{ij}^{GT} are the corresponding ground truth coordinates.

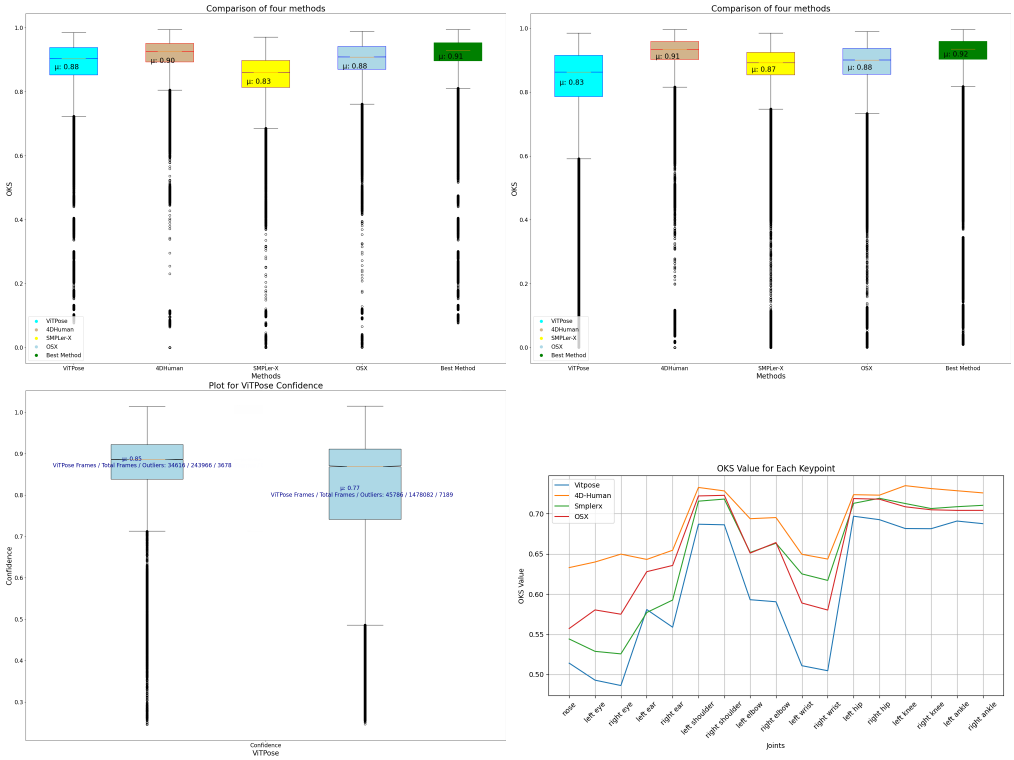

Fig. 3. Top: Comparison of different 2D Pose estimation methods: original videos (Left) and with chromatic adaptation settings 1 to 7 (Right); **Bottom**: ViTPose contribution to frames selected for best OKS (left); Visualization of OKS per joint with filtering (right).

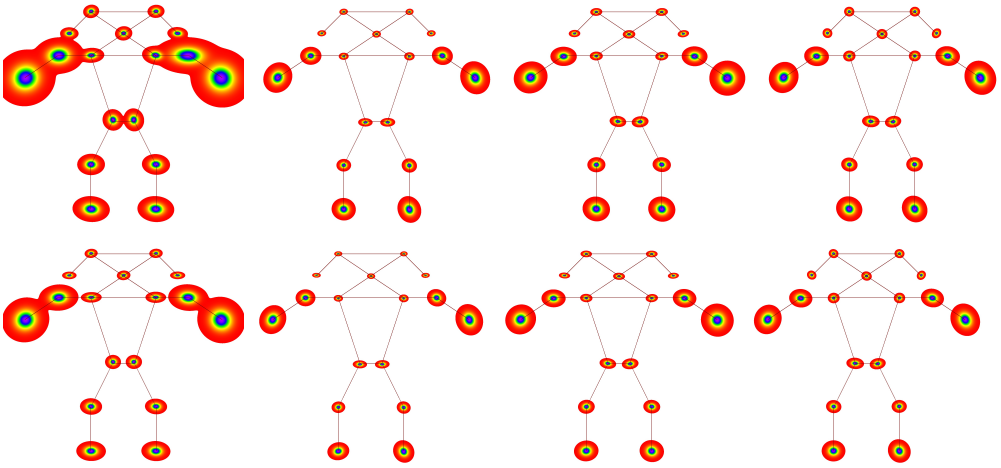

Fig. 4. Visualization of per joint stability of different methods without filtering (1st row) and with filtering (2nd row) : Vitpose, 4D Human, OSX and SmplerX (From left to right)

3.4 Visualization Methods

In the context of 2D human pose estimation in real world scenario, understanding the stability of individual joints is crucial for assessing the robustness and reliability of various pose estimation methods under different conditions. To address this, we adopt a methodological approach centered on a 2D Gaussian distribution to characterize the stability of each joint. This representation allows us to quantify the uncertainty or variability inherent in the estimated joint positions.

For each joint j, we compute the error between the predicted and ground truth coordinates. This error captures the deviations in both the horizontal (x) and vertical (y) directions. We computed the covariance matrix which encodes the relationship between errors in x and y dimensions, providing insights into the joint's stability in both directions. From the covariance matrix, we derive the standard deviations along the x and y axes. These standard deviations serve as measures of the joint's stability, indicating the extent of variability or dispersion in the estimated joint positions.

Utilizing the mean error and covariance matrix, we construct a multivariate normal distribution, representing the joint's positional uncertainty. By evaluating the Probability Density Function (PDF) of this distribution across a grid of coordinates spanning the image, we obtain a 2D Gaussian map. This map visualizes the likelihood of each joint being located at each point in the image, with higher probabilities indicating greater stability. The resulting 2D Gaussian maps provide valuable insights into the stability of various pose estimation methods under varying color illuminations. Additionally, the standard deviations along each axis offer quantitative measures of joint stability, facilitating comparisons across joints and pose estimation methods.

3.5 Choice of performances dataset

We initiated our study with the AIST++ dataset [23], comprising 12,670 videos that span a variety of dance genres, feature multiple dancers, that were captured from 9 distinct camera angles per performance, as highlighted in section 2.1. Our methodology involves processing these video frames to detect individuals, apply filtering, track, and interpolate to ensure a consistent temporal sequence across complex and rapid movements. This forms the foundation for our comparative analysis of various human body pose estimation techniques using data captured under optimal lighting conditions.

Subsequently, we introduced variations in illumination conditions by simulating color shifts using the chromatic adaptation method outlined in section 3.2. We randomly chose 50 dance sequences, ensuring each was represented by 9 different camera perspectives, same as original dataset so that there is no imbalance in the data due to skipped camera angles, totaling 450 videos. These were then augmented with 7 distinct color shifts settings (refer Table 1 (b)) previously described, yielding 3,150 videos tailored for assessing the precision and robustness of 2D pose estimation under varied illumination conditions. For both datasets —the original 12,670 videos and the 3,150 videos subjected to color illumination shift— we evaluated the performance of pose estimation models, both with and without the integration of supplementary filtering and interpolation algorithms. This comprehensive approach allows to gauge the effects of these additional processing on the accuracy and consistency of pose estimation, as well as to understand the influence of simulated lighting conditions.

4 Results and Discussions

4.1 Accuracy

Using the Object Keypoint Similarity (OKS) metric, we computed the average OKS across all estimated poses, for each pose estimation method, considering both the original videos and those with color illumination shift. Fig. 3(top) illustrates the OKS plot for each method. The best feature-based pose estimation method, Vitpose, exhibited strong performance, with an OKS of 0.88 on the original video dataset, comparable to the best shape-based method, 4D-Human, with an OKS of 0.9. However, when evaluating performance on the color illumination shift dataset, Vitpose's accuracy declined to 0.83, whereas the accuracy of shape-based methods remained consistent. Further analysis, illustrated in Fig. 3(bottom left), delved into the distribution of frames utilizing Vitpose versus the best shape-based method per frame within both datasets. Intriguingly, in the original video dataset, the ratio of frames employing Vitpose versus the best shape-based method was 0.142. However, this ratio diminished to 0.031 when analyzing the dataset with color illumination shift.

These findings underscore the susceptibility of feature-based pose estimation methods, such as Vitpose, to variations in color illumination. The observed decrease in accuracy and the shift in the ratio of method usage between the datasets underlines the potential challenges in deploying feature-based pose estimation methods in real-world scenarios with diverse lighting conditions.

Model	Metric	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12	J13	J14	J15	J16	J17
ViTPose	OKS	0.50	0.48	0.48	0.58	0.55	0.69	0.68	0.60	0.60	0.54	0.53	0.70	0.69	0.68	0.68	0.70	0.69
	MAE_x	7.73	7.89	8.01	9.35	10.56	10.36	10.58	17.53	18.14	20.86	21.63	**8.95**	**9.53**	9.45	9.56	10.13	10.39
	MAE_y	6.50	6.51	6.73	5.45	5.75	6.31	6.44	12.28	12.42	19.17	20.08	13.81	14.02	10.08	9.89	7.37	7.54
	SD_x	9.77	9.04	9.60	10.40	11.59	14.50	15.31	28.51	38.91	33.87	34.65	11.86	11.67	15.64	15.85	21.05	20.76
	SD_y	8.08	7.60	7.41	6.92	7.13	8.44	10.56	20.61	20.77	31.99	33.34	12.13	12.99	12.05	11.71	14.63	14.66
4D-Human	OKS	**0.64**	**0.65**	**0.65**	**0.65**	**0.66**	**0.74**	**0.74**	**0.70**	**0.70**	**0.66**	**0.65**	**0.73**	**0.73**	**0.74**	**0.74**	**0.74**	**0.73**
	MAE_x	**4.38**	**4.26**	**4.05**	**5.11**	**5.21**	**4.44**	**4.83**	**7.34**	**7.01**	**9.01**	**9.66**	10.27	10.64	**3.98**	**4.18**	**5.36**	**5.64**
	MAE_y	**3.69**	**3.14**	**3.03**	5.17	4.51	**5.11**	5.71	**6.48**	**6.39**	**9.31**	**9.94**	7.75	7.51	**4.97**	**5.68**	**5.98**	**6.80**
	SD_x	**4.88**	**4.80**	**4.52**	**5.02**	**4.97**	**5.59**	**5.89**	**12.03**	**11.80**	**16.34**	**16.88**	**8.14**	**8.43**	**8.28**	**8.61**	**13.66**	**13.42**
	SD_y	**3.90**	**3.50**	**3.54**	**3.43**	**3.49**	**4.04**	**4.58**	**10.15**	**10.39**	**17.18**	**18.61**	**4.70**	**4.87**	**7.11**	**8.00**	**12.68**	15.06
ViTPose(f)	OKS	0.51	0.49	0.49	0.58	0.56	0.69	0.69	0.59	0.59	0.51	0.50	0.70	0.69	0.68	0.68	0.69	0.69
	MAE_x	6.74	7.04	7.17	8.49	9.45	9.43	9.51	15.83	16.26	18.76	19.25	**7.97**	**8.53**	8.72	8.78	9.46	9.75
	MAE_y	5.96	6.00	6.23	5.05	5.33	5.80	5.93	10.69	10.77	17.07	17.67	13.03	13.28	9.37	9.21	6.61	6.81
	SD_x	7.68	7.36	7.57	8.13	8.54	11.66	11.54	22.25	22.60	26.72	27.02	9.01	9.19	12.47	12.46	16.68	16.39
	SD_y	5.42	5.16	5.21	4.25	4.30	5.93	6.32	14.85	15.03	25.34	26.25	7.98	7.95	9.30	8.84	11.22	11.34
4D-Human(f)	OKS	**0.63**	**0.64**	**0.65**	**0.64**	**0.65**	**0.73**	**0.73**	**0.69**	**0.70**	**0.65**	**0.64**	**0.72**	**0.72**	**0.73**	**0.73**	**0.73**	**0.73**
	MAE_x	**4.24**	**4.14**	**3.86**	**5.11**	**5.08**	**4.36**	**4.71**	**7.17**	**6.77**	**8.85**	**9.35**	10.30	10.56	**3.90**	**4.05**	**5.12**	**5.41**
	MAE_y	**3.57**	**3.06**	**2.93**	5.19	4.49	**5.10**	5.69	**6.28**	**6.18**	**9.17**	**9.68**	7.68	7.43	**4.75**	**5.50**	**5.53**	**6.38**
	SD_x	**4.63**	**4.58**	**4.29**	**4.90**	**4.80**	**5.28**	**5.53**	**11.21**	**10.90**	**15.33**	**15.67**	**8.08**	**8.25**	**7.75**	**7.88**	**12.40**	**12.04**
	SD_y	**3.15**	**2.81**	**2.80**	**2.67**	**2.84**	**3.76**	**4.39**	**9.54**	**9.82**	**16.31**	**17.68**	**4.51**	**4.70**	**6.49**	**6.96**	**10.19**	12.68

Table 2. Comparison of OKS, Mean Error (x,y) and Standard Deviation (x,y) for Models : VitPose, 4D-Human (without filter and with filtering) for Joints J1 to J17 (nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle) without filtering

Furthermore, considering per-joint accuracy under color illumination shift, as depicted in Fig. 3 (bottom right), the per-joint plot for all methods exhibited a similar and consistent pattern across all methods. Although each method's per-joint plot varied in vertical position on the OKS scale, indicating differing overall accuracies, they all experienced similar fluctuations in accuracy at corresponding joints. Notably, joints around the face (eyes, nose, and ears) consistently exhibited the lowest accuracy, while joints at the shoulders, hips, knees, and ankles demonstrated the highest accuracy. This observation underscores the challenges posed by color illumination shift on pose estimation accuracy, particularly in facial regions.

4.2 Stability

To analyse the stability of individual joints under color illumination shift, we employed our 2D Gaussian distribution visualization method (detailed in Section 3.4) to conduct both quantitative and qualitative analyses. Fig. 4 (top) and Table 2 showcase the stability of various pose estimation methods when our filtering technique is not applied.

From Fig. 4 (top), it is evident that the feature-based method, Vitpose, exhibits the highest instability across all joints when no filtering method is employed, while 4D-Human emerges as the most stable method. This qualitative observation is further supported by the quantitative analysis provided in Table 2. Specifically, 4D-Human demonstrates the lowest standard deviation along the horizontal axis across all joints, indicating superior stability in the horizontal direction. Additionally, 4D-Human boasts the lowest standard deviation along the vertical axis across all joints except for Joint 17 (right ankle). These findings highlight 4D-Human's exceptional stability across both horizontal and vertical axes under color illumination shift followed by OSX (refer supplementary material), which shows the best stability along the vertical axis for Joint 17. This stability trend reinforces the notion that shape-based methods consistently outperform feature-based methods over a temporal video when no filtering is applied.

The observed instability in feature-based methods such as Vitpose can be attributed to their reliance on local features within the image for detecting keypoints, which may lead to sensitivity to noise or errors in feature detection. Feature-based methods may lack sufficient global contextual information about the overall body pose, making them more susceptible to high noise or instability in pose estimation, particularly in challenging lighting conditions or noisy environments. This limitation becomes evident in cases of mismatched joint errors under low lighting, especially involving symmetrical joints such as left-right hips, knees, and ankles flipping with each other.Upon implementing our filtering technique to mitigate such noise, notable improvements in stability were observed across all methods. This refinement is depicted in Fig. 4 (bottom) and Table 2, where significant stability enhancements are evident for Vitpose, alongside increased stability across shape-based pose estimation methods.

5 Limitations and Future Works

One promising direction for future exploration involves the development of more sophisticated relighting techniques. Chromatic adaptation generally performs well for opaque

surfaces in human pose estimation, but its effectiveness diminishes for translucent and fluorescent materials, which can be part of the human subject such as clothing items. For example, under blue lighting, distinguishing facial features may become challenging while high-frequency details like pores and fine lines become more noticeable. Enhancing the capability to simulate diverse lighting conditions with greater accuracy and realism than chromatic adaptation, could help in more accurate experimentation and analysis, which will further enrich our understanding of pose estimation robustness under varying environmental factors.

In addition, there is a compelling opportunity to propose and explore novel 2D methods for pose estimation in challenging scenarios. We advocate for a rethinking of traditional approaches and propose a new method that combines the strengths of feature-based and shape-based techniques. This hybrid approach entails training a model to estimate pose keypoints while simultaneously capturing the contour and shape of the human body. By integrating these elements, the proposed method aims to deliver enhanced accuracy, stability, and reliability across varying environmental conditions such as low light.

Acknowledgments. This work was supported by HORIZON-CL2-2021-HERITAGE-01-04 grant. Project: 101061303 — PREMIERE [20].

References

1. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., et al.: Smpler-x: Scaling up expressive human pose and shape estimation (2023)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime Multi-person 2d Pose Estimation Using Part Affinity Fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Cheng, B., Xiao, B., Wang, J., Shi, H., et al.: Higherhrnet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
5. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., et al.: Strongsort: Make deepsort great again. IEEE Transactions on Multimedia (2023)
6. Duay, K.: Players detection and tracking for eSports videos analysis with dataset generation from minimap. Master's thesis, Itä-Suomen yliopisto (2023)
7. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–17 (2022)
8. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023)
9. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (10 2017)
10. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
11. Khirodkar, R., Chari, V., Agrawal, A., Tyagi, A.: Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
12. Lam, K.M.: Metamerism and color constancy. Ph. D. Thesis, University of Bradford (1985)

13. Li, J., Wang, C., Zhu, H., Mao, Y., et al.: Crowdpose: Efficient Crowded Scenes Pose Estimation and a New Benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (6 2019)
14. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21159–21168 (2023)
15. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312
16. Loper, M., Mahmood, N., Romero, J., et al.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (2015)
17. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs] (2016), http://arxiv.org/abs/1603.00831, arXiv: 1603.00831
18. Ojeleye, J.O.: Developing tools and methods for human body pose estimation. Master's thesis, Itä-Suomen yliopisto (2023)
19. Patel, P., Huang, C.H.P., Tesch, J., et al.: AGORA: Avatars in geography optimized for regression analysis. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2021)
20. Premiere: Premiere - Performing arts in a new era (3 2024), https://premiere-project.eu/
21. Qiu, L., Zhang, X., Li, Y., Li, G., Wu, X., Xiong, Z., et al.: Peeking into Occluded Joints: A Novel Framework for Crowd Pose Estimation, pp. 488–504. Springer (2020)
22. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
23. Tsuchida, S., Fukayama, S., Hamasaki, M., et al.: Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In: Proc. of the 20th Int. Society for Music Information Retrieval Conf., ISMIR. pp. 501–510 (2019)
24. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
25. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines (2016)
26. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)
27. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose++: Vision transformer foundation model for generic body pose estimation. arXiv preprint arXiv:2212.04246 (2022)
28. Zhang, Y., Sun, P., Jiang, Y., Yu, D., et al.: Bytetrack: Multi-object tracking by associating every detection box (2022)