Analysis of Human Motion Based on AI Technologies: Applications for Safeguarding Folk Dance Performances

Nikos Grammalidis, Iris Kico, and Fotis Liarokapis

Abstract Analysis of human motion is an important research area in computer vision with numerous applications. Recent projects, such as EU i-Treasures and TERPSICHORE projects conduct research in this field to improve the capture, analysis and presentation of Intangible Cultural Heritage (ICH) using ICT-based approaches. The final goal is to document these forms of intangible heritage and to capture the associated knowledge in order to safeguard and transmit this information to the next generations. In addition, these approaches can give rise to new services for research, education and cultural tourism. They can also be used by creative industries (e.g. companies performing film, video, TV or VR applications production), as well as by local communities, creating new local development opportunities by promoting local heritage. This paper first reviews some very recent state of the art approaches based on deep learning which can achieve impressive results in recovering human motion (2D or 3D) and structure (skeleton with joints or realistic 3D model of the human body). Based on such approaches, we then propose a dance analysis approach, currently under development in TERPSICHORE project. Preliminary results are presented and, finally, some conclusions are drawn.

Keywords Human motion analysis from video · Artificial intelligence · Deep learning · Folk dance preservation · Applications

N. Grammalidis (✉)
Information Technologies Institute, CERTH, Thessaloniki, Greece
e-mail: ngramm@iti.gr

I. Kico · F. Liarokapis
Faculty of Informatics, Masaryk University, Brno, Czech Republic
e-mail: iriskico@mail.muni.cz; liarokap@fi.muni.cz

A. Kavoura et al. (eds.), *Strategic Innovative Marketing and Tourism*,
Springer Proceedings in Business and Economics,
https://doi.org/10.1007/978-3-030-36126-6_35

1 Introduction

Analysis of human motion is an important research area in computer vision with numerous applications. In the general case, analysing human motion from video can be a complicated problem, considering cluttered backgrounds, illumination variations, occlusions, self-occlusions, different clothing and multiple moving objects. In the past, many over-simplifying assumptions were often made to tackle these ill-posed problems or certain constraints were imposed. For instance, scene segmentation is simplified by assuming a moving person in front of a static background. However, recently, great advances were made in the field, mainly due to the efficiency of deep learning techniques, and particularly the Convolutional Neural Networks (CNN), a class of deep neural networks most applied to analysing visual imagery. Leveraging on the availability of big data and increased GPU computational efficiency, real-time methods to estimate multiple human motion with great accuracy were developed.

Dance performances, either as an autonomous form of art and expression, or as a part of the music and/or sound culture, were always important for human societies through the ages. Dances convey different messages according to the context, and focus on aesthetics or artistic aspects (contemporary dance, ballet dance), the cultural and social aspects (folk dances, traditional dances), storytelling (symbolic dances) or spiritual meanings (whirling dervishes). Especially folk dances are strongly linked to local identity and culture. The know-how of these dances survives at the local level through small groups of people who gather to learn, practice and preserve these traditional dances. Therefore, there is always the risk that certain elements of this important form of intangible cultural heritage could die out or disappear if they are not safeguarded and transmitted to the next generation. Therefore, their preservation for the next generations is of major importance.

In this paper, a number of some recent state-of-the-art approaches in the areas of analysis of human motion using optical and IR sensors are reviewed. Some of these techniques are currently tested within the framework of TERPSICHORE project, focusing on folk dance digitisation, analysis and its applications. The paper also proposes a new approach for automated dance choreography extraction from video, based on 3D skeleton joint extraction using deep learning. Such techniques offer multiple advantages and economic benefits to education, tourism, creative industries and cultural institutions.

The rest of this paper is structured as follows: In Sect. 2 presents a review some very recent state of the art approaches based on deep learning, which can achieve impressive results in recovering human motion (both 2D or 3D) and shape (realistic 3D model of the human body). In Sect. 3, we review some similar applications for dance analysis and visualization and propose an approach, currently under development in TERPSICHORE project, to estimate the choreography from videos in the wild. In Sect. 4, some indicative results are presented in dance videos captured during TERPSICHORE project, while in Sect. 5 some conclusions are drawn.

2 Human Motion Analysis Based on Deep Learning

Human body motion analysis and action recognition are two crucial tasks for understanding human behaviour and can be used for many different applications, including surveillance, human computer interaction, educational applications, games and many more. Pose estimation refers to the process of estimating the configuration of the underlying kinematic or skeletal articulation structure of a person [1]. Estimating human pose from video input is an increasingly active research area in computer vision that could give rise to numerous real-world applications, including dance analysis. Traditional methods for pose estimation model structures of body parts, mainly based on handcrafted features. However, such methods may not perform well in many cases, especially when dealing with occlusions on body parts.

Recently, great technological advances were made in 2D human pose estimation from simple RGB images, mainly due to the efficiency of deep learning techniques, and particularly the Convolutional Neural Networks (CNN), a class of deep neural networks most applied to analyzing visual imagery. A new benchmark dataset is introduced by Andriluka et al. [2], followed by a detailed analysis of leading human pose estimation approaches providing insights for the success and failures of each method. Some very effective open source packages have become increasingly popular, such as OpenPose [3], a real-time method to estimate multiple human poses efficiently developed at Robotics Institute of Carnegie Mellon University. OpenPose represents a real-time system to jointly detect human body, hand and facial keypoints (130 keypoints in total) on single images, based on Convolutional Neural Networks (CNN). More specifically, OpenPose extends the "Convolutional pose" approach proposed in [4] and estimates 2D joint locations in three steps: (a) by detecting confidence maps for each human body part, (b) by detecting part affinity fields that encode part-to-part associations and (c) by using a greedy parsing algorithm to produce the final body poses. In addition, the system computational performance on body keypoint estimation is invariant to the number of people detected in the image [3, 5].

In [6], a weakly supervised transfer learning method is proposed for 3D human pose estimation in the wild. It uses mixed 2D and 3D labels in a unified deep neural network that has a two-stage cascaded structure. The module combines a) a 2D pose estimation module, namely the hourglass network architecture [7], producing low-resolution heat-maps for each joint and b) a depth regression module, estimating a depth value for each joint. An obvious advantage from combining these modules in a unified architecture is that training is end-to-end and fully exploits the correlation between the 2D pose and depth estimation sub-tasks. Furthermore, in [8], a real-time method is presented to capture the full global 3D skeletal pose of a human using a single RGB camera. The method combines a CNN-based pose regressor with a real-time kinematic skeleton fitting method, using the CNN output to yield temporally stable 3D global pose reconstructions based on a coherent kinematic skeleton. The authors claim that their approach has comparable (and in some cases better) performance than Kinect and is more broadly applicable than RGB-D solutions (e.g. in outdoor scenes or when using low-quality cameras). RGB-D (Red, Green, Blue plus Depth) cameras provide per-pixel depth information aligned with image pixels from a standard camera. In [9], a fully feedforward CNN-based approach is proposed for monocular 3D human pose estimation from a single image taken in an uncontrolled environment. Authors use transfer learning to leverage the highly relevant mid- and high-level features learned on the readily available in-the-wild 2D pose datasets in conjunction with the existing annotated 3D pose datasets. Furthermore, a new dataset of real humans with ground truth 3D annotations from a state-of-the-art markerless motion capture system is produced.

A promising recent advancement is the recovery of parameterized 3D human body surface models, instead of simple skeleton models. This paves the way for a broad range of new applications, such as foreground and part segmentation, avatar animation, virtual reality (VR) applications and many more. In [10], dense human pose estimation is performed by mapping all human pixels of an RGB image to a surface-based representation of the human body. The work is inspired by the DenseReg framework [11], where CNNs were trained to establish dense correspondences between a 3D model and images ‘in the wild’ (mainly for human faces). The approach is combined with the state-of-the art Mask-RCNN system [12], resulting to a trained model that can efficiently recover highly accurate correspondence fields for complex scenes involving tens of persons with moderate computational complexity. In [13], a “Human Mesh Recovery” framework is presented for reconstructing a full 3D mesh of a human body from a single RGB image. Specifically, a generative human body model, SMPL [14] is used, which parameterizes the mesh by 3D joint angles and a low-dimensional linear shape space. The method is trained using large-scale 2D keypoint annotations of in-the-wild images. Convolutional features of each image are sent to an iterative 3D regression module, whose objective is to infer the 3D human body and the camera in a way that its 3D joints project onto the annotated 2D joints. To deal with ambiguities, the estimated parameters are sent to a discriminator network, whose task is to determine if the 3D parameters correspond to bodies of real humans or not. The method runs in real-time performance given a bounding box containing the person. Additional information and reviews of the progress in the field can be found in recent literature [15–17].

3 Applications for Dance Analysis and Visualisation

These rapid advances have already started to affect numerous fields, including dance analysis and visualization. In [18], Chen et al. employ a powerful deep learning architecture, namely a Generative Adeversarial Network, GAN, to transfer a dancing performance to a novel (amateur) target after only a few minutes of the target subject performing standard moves. Using pose detections as an intermediate representation between source and target, we learn a mapping from pose images to a target subject’s appearance. In other words, new realistic dance sequences of the amateur target, which are fun but also important for learning purposes, are generated from a 2D skeleton sequence extracted from a video of an expert. To extract pose keypoints for the body, face, and hands, they use the OpenPose state of the art pose detector [3], while two additional improvements are introduced: (a) improved temporal smoothness of the generated videos is achieved by conditioning the prediction at each frame on that of the previous time step and (b) the facial realism of the results is improved by including a specialized GAN trained to generate the target person's face. For the image translation stage of our pipeline, the architectures proposed by Wang et al. in the pix2pixHD model [19] are used.

In [20], a system that predicts 3D positions using given only 2D joint locations is proposed. Using a state-of the-art 2D detector, a relatively simple deep feedforward network, and training using 3D positions from Human3.6 M dataset, very promising results are obtained for estimating 3D from images in the wild.

However, an even more promising approach that extends the previous approach is presented in [21]. This approach also estimates 3D joint locations from corresponding 2D locations, but 3D information is used only implicitly, as a GAN network is used to estimate the z-values for each joint, so that the resulting 2D joint projections match those provided in the input (up to a rotation θ that can be estimated). Interesting improvements introduced by the authors are:

(a) the assumption of a simplified horizontal camera model, as generally the camera model is not available from images in the wild and
(b) the introduction of a new heuristic loss function, based on a simple body constraint (right shoulder is always to the right side of the face) to discard "wrong/inverted" camera poses.

In this chapter, we attempt to leverage the powerful 3D joint location detector proposed in [21], combined with a robust 2D joint detector, such as [3] or [22] for dance analysis of traditional folk dances. The final aim is the automated extraction of choreography from any video sequence or movie in the wild.

The choreography, which is the most basic element of dance, can be represented using special symbols to express the body configuration and movement of each body part with respect to time.

Dance Notation systems, describing the dance using symbols, appeared for the first time ever since the fifteenth century, and to date there are over eighty [23], although few of them are used in the modern era. Common dance notation systems include Labanotation [24], DanceWriting and others. Labanotation, the most widely used of these systems allows the recording and representation of any choreography and generally human movement. It was first proposed by the dancer and theorist Rudolf Laban in 1928. The motion analysis is based on the concepts of space, anatomy and dynamics of movement. It uses abstract symbols to describe the movement, providing a well-structured language with rich vocabulary and clear semantics, based on Laban Movement Analysis (LMA). The LMA defines four basic traits of movement: body, effort, shape, space (as well as two subordinates: relationship and expression). In [25], an innovative choreography generation system is presented, namely *chor-rnn*, that can generate novel choreographic material in the nuanced choreographic language and style of an individual choreographer. Chor-rnn is a deep Recurrent Neural Network (RNN) trained on raw motion capture data that can generate new dance sequences for a solo dancer. It can also be used for collaborative human-machine choreography or as a creative catalyst, to provide inspiration for a choreographer.

However, in the specific case of traditional folk dances, the choreography is usually simple and periodic, consisting of a set of simple steps, so only the movement of the lower part of the body is important. In i-Treasures project [26], we used a simplified notation system to describe each individual step in a folk dance period. Specifically, for each individual step of the period of the dance has the following features:

- StepTimestamp: timestamp when the step starts
- stepFoot: Foot that moves (Left, L or Right, R), so the other foot is assumed to be supporting the body, i.e. the 2D ankle locations remain fixed. Note that 3D locations cannot be used in this case, as these are always relative to the center joint (hip).
- StepDirection: Direction of foot movement (Left, L or Right, R). The direction of movement can be predicted by examining the variation of the magnitude of the vector v (between the left and right 3D ankle joints
- StepLift: whether the foot is lifted from the ground or not (e.g. in Tsamiko dance, men dance by lifting the foot, while women do not lift the foot)

Within i-Treasures project, we attempted to estimate this information by examining motion capture data (e.g. from VICON or Kinect), however the use of standard uncalibrated cameras opens exciting new opportunities as the approach can be applied to dance videos in the wild (e.g. from Youtube).

4 Experimental Results and Discussion

The proposed approach was tested using a set of recordings of various Greek folk dances, captured for the needs of the TERPSICHORE project. Specifically, input video from two synchronized cameras (front and profile views) was used. In Fig. 1, results of the 3D estimation of joints using the approach [21] are illustrated. Results for the proposed choreography analysis approach will be presented in the conference.

Furthermore, we applied [13] to predict and visualize the 2D joint movements and a full 3D mesh of a parameterised human body model from a single RGB image. Some predictions were accurate (Fig. 2), however some instabilities were observed, due to the large number of parameters to be estimated (85 parameters in total). However, we believe that these results can be significantly improved, e.g. by properly constraining some of these parameters, e.g. camera calibration if the camera is static or known.

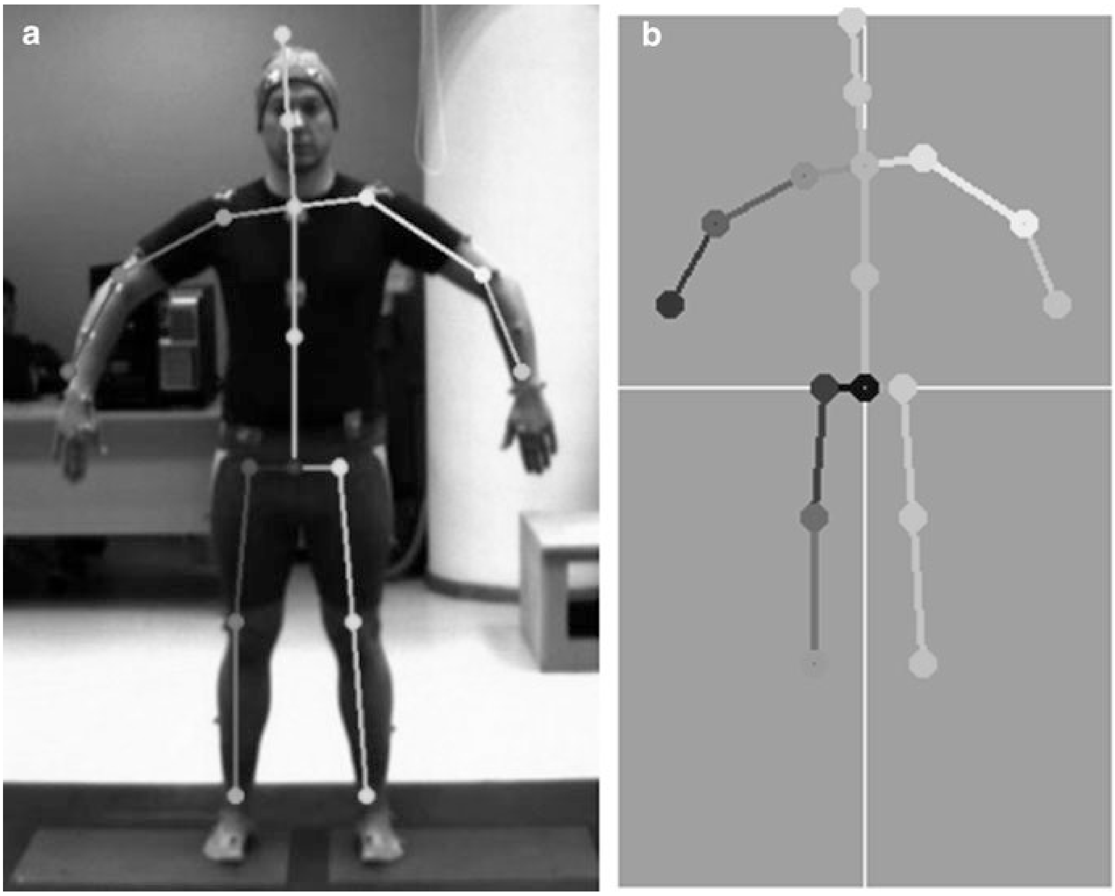

Fig. 1 2D human joint estimation using OpenPose and (b) corresponding estimated 3D joint locations using [21]

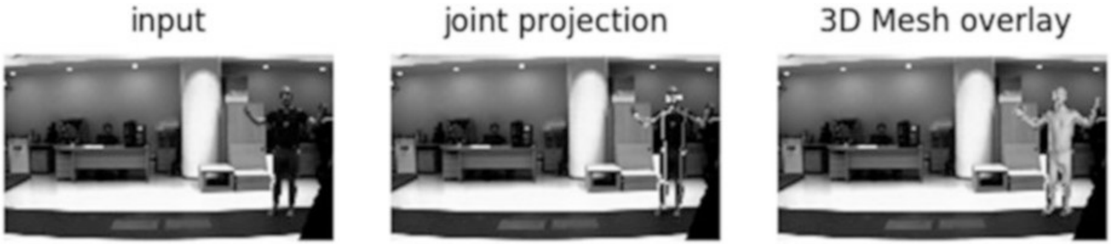

Fig. 2 Sample input frame, estimated 2D joint locations and 3D mesh overlay

5 Conclusions

Modern deep learning techniques have resulted to major research advances in 2D and 3D human pose estimation, which offer great advantages and give rise to new exciting applications in many fields, including applications for automated dance analysis from videos in the wild.

References

1. Moeslund TB, Hilton A, Kruger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Und 104:90–126. https://doi.org/10.1016/j.cviu.2006.08.002
2. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), Columbus, OH, USA, June 23–28, 2014. IEEE, Columbus, OH
3. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: The IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, July 21–26, 2017. IEEE, Honolulu, HI
4. Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, June 27–30, 2016. IEEE, Las Vegas, NV, pp 4724–4732
5. Simon T, Joo H, Matthews I, Sheikh Y (2017) Hand keypoint detection in single images using multiview bootstrapping. In: The IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, July 21–26, 2017, vol 2. IEEE, Honolulu, HI
6. Zhou X, Huang Q, Sun X, Xue X, Wei Y (2017) Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: 2017 IEEE international conference on computer vision (ICCV), Venice, Italy, October 22–29, 2017. IEEE, Venice
7. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: Liebe B, Matas J, Sebe N, Welling M (eds) Proceedings of 14th European conference, European conference on computer vision (ECCV) 2016, Amsterdam, The Netherlands, October 11–14, 2016. Springer, Cham, pp 483–499
8. Mehta D, Sridhar S, Sotnychenko O, Rhodin H, Shafiei M, Seidel HP et al (2017) Real-time 3D human pose estimation with a single RGB camera. ACM Trans Graph 36(4):44. https://doi.org/10.1145/3072959.3073596
9. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, Theobalt C (2017) Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 2017 International conference on 3D vision (3DV), Qingdao, China, October 10–12, 2017. IEEE, Qingdao, pp 506–516
10. Güler RA, Neverova N, Kokkinos I (2018) DensePose: dense human pose estimation in the wild. In: Proc. CVPR
11. Güler RA, Trigeorgis G, Antonakos E, Snape P, Zafeiriou S, Kokkinos I (2017) DenseReg: fully convolutional dense shape regression in-the-wild. In: The IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, July 21–26, 2017. IEEE, Honolulu, HI
12. He K, Gkioxari G, Dollar P, Girshick R, Mask R-CNN (2017) Proceedings of IEEE international conference on computer vision (ICCV), Venice, Italy, October 22–29, 2017. IEEE, Venice
13. Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 7122–7131
14. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: a skinned multi-person linear model. ACM Trans Graph 34(6):248:1–248:16, https://doi.org/10.1145/2816795.2818013
15. Gong W, Zhang X, Gonzalez J, Sobral A, Bouwmans T, Tu C, Zahzah E (2016) Human pose estimation from monocular images: a comprehensive survey. Sensors 16(12):1996. https://doi.org/10.3390/s16121966
16. Ke S, Thuc HLU, Lee YJ, Hwang JN, Yoo JH, Choi KH (2013) A review on video-based human activity recognition. Computers 2(2):88–131. https://doi.org/10.3390/computers2020088

17. Neverova N (2016) Deep Learning for Human Motion Analysis. PhD Thesis. Universite de Lyon, Lyon. https://doi.org/10.13140/RG.2.1.1255.8961
18. Chan C, Ginosar S, Zhou T, Efros AA (2018) Everybody dance now. arXiv preprint arXiv:1808.07371
19. Wang T-C, Liu M-Y, Zhu J-Y, Tao A, Kautz J, Catanzaro B (2017) High-resolution image synthesis and semantic manipulation with conditional GANs. arXiv preprint arXiv:1711.11585
20. Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3d human pose estimation. In: ICCV
21. Yasunori K, Ogaki K, Matsui Y, Odagiri Y (2018) Unsupervised adversarial learning of 3D human pose from 2D joint locations. arXiv preprint arXiv:1803.08244
22. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer, Cham, pp 483–499
23. Hutchinson-Guest AD (1989) Choreo-graphics: a comparison of dance notation systems from 15th century to the present. In: International conference exploring research. Routledge, New York, p 194
24. Laban R (1928) Schrifttanz. Universal, Wein
25. Crnkovic-Friis L (2016) Generative choreography using deep learning. In: 7th International conference on computational creativity, ICCC2016
26. Dimitropoulos K, Tsalakanidou F, Nikolopoulos S, Kompatsiaris I, Grammalidis N, Manitsaris S, Hadjileontiadis L (2018) A multimodal approach for the safeguarding and transmission of intangible cultural heritage: the case of i-treasures. IEEE Intell Syst 33(6):3–16