2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

Towards a Humanoid-Oriented Movement Writing

Adrian Stoica, Hyung Ju Suh, Steven M. Hewitt, Sarah Bechtle, Anna Gruebler and Yumi Iwashita

***Abstract*— This paper introduces humanoid-oriented movement writing (HOM), focusing on a notation in which body postures allow easy visual interpretation by both humans and humanoid robots. HOM Writing, derived from Sutton Movement Writing and Shorthand [1], is a natural modality for encoding the movements that humans perform during various work activities. Beyond its use as a record of human movement, the intent is to use the writing as a modality of communicating to robots what movements to execute. Humanoid robots could directly map the key postures represented in the notation to their own postures, imitating the postures captured in the description, and calculating intermediate postures by interpolation. A motion generator would ensure the motor control needed to create continuous movements. This paper focuses on the generation of the activity movement scripts, and addresses two modalities of producing the scripts: 1) 'hand-coded' by a human, using an editor, and 2) automatic extraction from video, e.g. from video-recordings of a human performing an activity. A software tool developed to allow easy script writing, the HOM Editor, is described and illustrated in hand-coding of a sequence of movements. The automatic generation of scripts from video is done using the pose estimation system by Yang and Ramanan [2], which takes an image and produces the joint coordinates of the limb parts; this is illustrated with a task of moving and arranging chairs. The posture extraction from a video provided by a single camera may often lead to occlusions of body parts during activities in which objects are manipulated. We show the advantages of using an additional camera, which significantly increases the correct posture estimation, and discuss how to further improve the automatic generation of scripts.**

I. Introduction

Cooperation between humans and robots is increasingly important. The behaviors, actions and activities that robots are instructed to perform are becoming more complex. At high level these are specified in natural language, yet, the exact motions a robot needs to execute are most likely specified in some programming language. In order to represent and communicate movements there is a growing need of an efficient notation system for describing robot and human motion, one that can be understood by robots, and ideally is natural and intuitive to the humans. Schematic notations and short hands have been developed in all domains that involve body movement, from dance to martial arts, with different degrees of attention to the accuracy of the postures. The movement writings are often complemented by language-based descriptions that accompany them (e.g. in martial arts - 'quickly raise the hands above the head then use left fist to strike to the right'). Description in natural language clarifies the drawings by providing further information, such as tempo and intensity.

Learning from human instruction is a topic that has recently attracted a lot of interest, as indeed it is easier to show a movement and have the robot imitate it, or replicate it by analogy, then it is to program a robot. Unfortunately, the big gap between these two extremes (programming and demonstration) does not have much in techniques to cover the space. If one opts for robot programming, all low-level descriptions are very platform-specific, sacrificing portability. If one opts for human demonstration, unless we use video/photo-recordings, there is no good modality to store, to document, or to transmit the motion. This disadvantage points to a fertile ground for a movement writing system, as a tool to document movements, actions, activities demonstrated by a human, or simply performed by humans in the course of executing daily routine motions (without an emphasis on demonstration and teaching) in a fluid and organic way.

This paper introduces humanoid-oriented movement writing (HOM) for encoding human movements executed during various work activities, to be then transferred to robots for execution. Fig. 1 illustrates the most popular movement writing systems, i.e. the Laban notation [3], the Benesh notation [4], and the Sutton Movement Writing and Shorthand [1]. The paper is focusing on a notation in which body postures allow easy visual interpretation, by both humans and humanoid robots, thus using the Sutton notation as an inspiration.

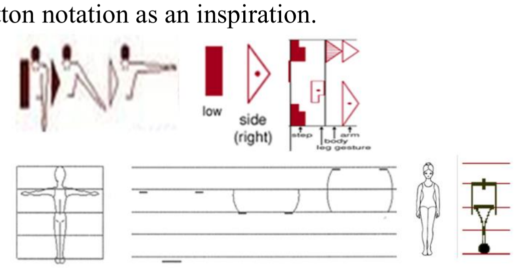

Figure 1. Movement writing in dance/choreography – (top) Laban notation; (bottom left) Benesh notation, (bottom righ) Sutton notation. Laban and Benesh systems appear overly abstract and non-intuitive to the non-specialist. Sutton offers a more intuitive system for imitation by humanoid robots, in terms of replicating human postures.

Humanoid robots would directly map the key postures represented in the notation to their own, imitating them, and also calculating the intermediate postures between them (key postures are equivalent to waypoints in robot navigation). A motion generator will ensure the motor control needed to create continuous movement. Although we envision later on to use

*Research performed at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA 91107, under a contract with the National Aeronautics and Space Administration (NASA) A. Stoica (e-mail: adrian.stoica@ jpl.nasa.gov) and Yumi Iwashita are with the Jet Propulsion Laboratory.

Hyung Ju Suh, a Caltech student, and Sarah Bechtle, are affiliated interns at JPL. S. M. Hewitt was an 2016 intern at the JPL, and is now with UC Berkeley. Anna Gruebler was a JPL visiting researcher in 2013, while PhD student at U. of Tsukuba, Japan.

HOM to control humanoid robots, in this paper we focus on the generation of the activity movement scripts, and address two modalities of producing the scripts: 1) 'hand-coded' by a human using an editor, and 2) the automatic extraction of the movement scripts from video recordings of a human performing an activity. The HOM Editor, a software tool for easy script writing and movement editing, is described and illustrated. The generation of scripts automatically from video is done using the pose estimation system by Yang and Ramanan [2], which takes an image and produces the joint coordinates of the limb parts. The automation of the script generation process is illustrated with a basic task of moving and arranging chairs. The direct application of the video extraction from a single camera expectedly leads to occlusions of body parts during activities including objects to be manipulated. We show that the use of an additional camera significantly increases the correct posture estimation, and discuss how to further improve the automatic generation of scripts. The paper is organized as follows: Section II presents previous work related to movement writing system in robotic applications. Section III introduces in detail the HOM writing system. In Section IV and section V the editing tool and the automatic extraction of movement scripts from video are explained respectively. In section VI and VII our results are presented and discussed.

II. Previous Work

The idea of using movement writing for robotics has been proposed at least as early as 1998 [5], [6] and more work followed. The two most popular movement writing systems in dance and choreography are Laban [3] and Benesh [4]. Both could make powerful robot movement notation systems given the fact that both provide a complete language for movement description. The majority of prior work [7], [8], [9], [10] attempts to use the Laban notation for motion writing in robotics. The Laban system, which can be seen from the top figure in Figure 1, does not allow a direct reproduction of the movement from its pure observation; to the human eye the notation is less intuitive than the Sutton notation (Figure 1, bottom right). The same is true for the Benesh notation (Figure 1, bottom left). Laban and Benesh are not visually intuitive for interpretation by those not familiar with the encoding. We believe that humanoid robots would benefit more from a notation that describes postures in a more human -like and intuitive manner, as the Sutton notation does. In the following we will give an overview of previous work that used movement writing system in robotic applications.

In [7], Knight and Simmons adapt the Laban notation to the movement of a 2-DOF Aldebaran Nao[1] head and a 4 DOF Keepon[2]. The robot movement features were created manually, and no automatic movement detection was used. The study focused on determining how well users of Amazon Mechanical Turk[3] would correctly interpret the robot movement. They could achieve statistically significant results claiming that the used robots could convey complex expression to people when using the Laban notation. Even though the results were satisfactory from an interaction point of view, the robot movement was created manually by a human. Furthermore, this study only focused on a 4-DOF system, as the full body movement of a humanoid was not considered.

Samadani et al [8] describe in their paper how they adapted two existing quantification approaches of Laban components for hand and arm movements only. Six hand and arm motion paths were designed to convey six basic emotions. The hand and arm movements were Laban annotated by a "certified movement analyst (CMA)" to compute the statistical correlation between the CMA-annotated and the quantified Laban components. The results show that the correlation between the CMA annotated and the quantified outcome is high (~80%). In [9] Hachimura et al. again compared the Laban movement notation extracted algorithmically from a motion capturing system with the results of the analysis of a specialist. They achieve partially satisfactory results from the comparison but claim that a numerical formulation of the Laban movement notation is possible.

In [10] the authors present a framework for emotion recognition from video. The hand of a person was tracked, and the analysis of the tracking showed that acceleration and frequency characteristics of the hand are relevant to recognizing emotion. The authors argue that a computer encoding of the Laban movement can serve as a common language for expressing and interpreting emotional movements between robots and humans. The work presented in [11] and [12] formulates solutions for retargeting human motion to humanoid motion, which is of interest when adapting any movement writing system to robotic motion. In [11] human motion was captured by a motion capture system and then converted to humanoid movement. In [12], the human motion data was obtained from a human motion database and a pose tracking system, but only the human upper body motion was retargeted to the robot.

These previous attempts to retarget human motion to humanoid motion provide evidence to the interest in a robot imitating or following human movement. Nevertheless, the work in [11] and [12] stay short the formalization of a notation system that describes movement in a more generalized way. The majority of the previous work focuses on the Laban notation, and rarely considers the full anthropomorphic body movement. In this paper we propose an intuitive notation, derived from Sutton movement writing, modified to better serve humanoid robots. Additionally, we emphasize the use of movement / posture detection from video in contrast to motion capturing technology.

III. HOM Writing

The Sutton writing notation represents human postures through an abstract anthropomorphic stick figure. Stick figures are easy to visualize and interpret. Additionally, longer scripts contain complicated figures for notating the type of motion between poses. Poses often have accompanying symbols that describe the type of motion to be executed, the timeframes to

[1] Aldebaran Nao is a 58 cm tall humanoid robot with 25 DOF https://www.ald.softbankrobotics.com/en/cool-robots/nao

[2] Keepon is a 4 DOF robot toy http://www.beatbots.net/keepon-pro

[3] https://www.mturk.com/mturk/welcome

accomplish the transition between poses, and other detailed information that is less meaningful to the untrained eye. The descriptive power comes at the price of complexity. We derived HOM from the Sutton notation, modifying the stick figures to allow only straight segments, corresponding to robot links, as well as linearized rigid bone formations. It is developed to formally represent movement primitives and their composition in activities in a way that is easily recognizable by both humans and by robots using pure vision. To retain a complete language, HOM inherits all the conventions of the Sutton Writing notation, yet, further adaptation to suit robot bodies and robotic activities is needed. HOM includes the following elements visible in Figure 2 and 4:

1. *A sequence of poses is separated by vertical lines, similar to measures in Western music notation*

2. *All poses contain at least one actor as well as zero or more objects and tools that the actor can interact with*

3. *The actor is represented by eleven line segments, which indicate: The head (one line), the shoulders (one line), the spine (one line), the upper arms (two lines, one line per arm, the forearms (two lines, one line per arm), the hips (one line), the thighs (two lines, one line per leg), the lower leg (two lines, one line per leg)*

4. *The horizontal lines represent the normal position for the foot line, the knee line, the hip line and the shoulder line. The most upper line represents the height of the arm when lifted over the head.*

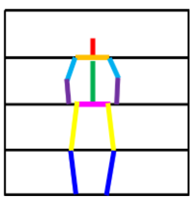

Figure 2. The HOM notation, colored to highlight different parts.

IV. Editing Tool and Example

An editor tool was developed to conveniently express the possible poses and motions in HOM. The editor allows easier writing of motion scripts, through manipulation of the angles between each limb segment. The expressions of the skeletons in terms of angles allow the skeleton to be properly normalized, regardless of the length of the limbs specific to an individual human, or robot. The software was developed as a multi-module flexible system using “toolkits” that can help express the motion writing notation. These toolkits not only allow the user to easily manipulate individual frames, but also introduce new features to the scripts. The features include functionalities such as saving and loading scripts, describing multiple users within the scripts, specifying exact time coordinates of each skeleton, and adding verbal descriptions along with the script. Figure 3 shows the editor defining five poses. The direction in which the skeleton faces, and the values of the joint angles are entered manually, and define the script shown in Figure 4.

facing	0.7	1.3	1.3	1.2	1.2
armAngleL	-20	-10	0	0	-50
armAngleR	-20	-10	0	0	-50
elbowAngleL	-50	-70	-10	-10	-30
elbowAngleR	-50	-60	-10	-10	-30
legAngleL	-10	-10	0	-30	-10
legAngleR	-10	-20	0	-30	-10
kneeAngleL	20	20	0	50	20
kneeAngleR	20	10	0	50	20
backBend	-30	0	0	-70	0
neckAngleV	0	0	0	0	0
neckAngleH	0	0	0	0	0

Figure 3. The input parameter angles in the editing tool.

Figure 4. The sequence of posture descriptions describing a motion

V. Automatic Extraction From Video

A. Extraction Method

1) Extraction of Joint Coordinates from Videos

For the extraction of joints, we used the software from Yang and Ramanan [2], which has shown relatively high *probability of a correct pose* (PCP) scores, while correctly preserving local rigidity on each limb. The software takes an image and produces the joint coordinates of the parts of the limb. The algorithm utilizes a pictorial structure framework [13], in which the parts of each limb are detected, then inferred using the relational constraints between each limb part to create a coherent skeleton structure. The skeleton is then augmented through various rotated and lengthened versions of the original configuration of the limbs, which provides better accuracy. The pose is then recognized through supervised learning on human pose databases, such as the Image Parse dataset [14]. For our purposes we extended the algorithm to input and output videos instead of single images. The detection software provides 26 positions of the joints in the detected human body per frame.

2) Selection of Motion Primitives

In order for the robot to perform a full movement, characteristic key postures have to be chosen in a way that it becomes possible to interpolate between two consecutive key postures and create a fluid movement. An attempt to automatically detect key postures from a video was made by choosing the points where the derivative of the joint positions is zero, i.e. the frames at which the velocity of a particular joint is near-zero. This resulted to be not trivial, as will be described in section VI.

3) Notation Conversion and Conformal Normalization

For a more general notation of the detected pose, we converted the 26-joint posture detected by the pose estimation software to our HOM notation. This was done by selecting the relevant 11 joints (see Figure 3) from the outputted 26-joint skeleton. Clearly some information is lost during the conversion, as not all the detected joints are considered. Yet, we believe that the representation is accurate enough, given the fact that the HOM notation attempts to provide a more generalized and normalized notation of the human posture, independent of the individual human or robot model. After detecting the relevant joints, the angles between the joints are calculated and imported to the editing software. Thus, we achieve conformal normalization of the pose (i.e. normalization of limb lengths while preserving the angles between the poses). The overall process is illustrated by Figure 5.

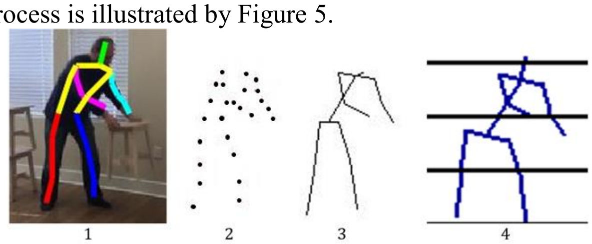

Figure 5. An illustration of the process used: (1) original detected pose, (2) detected 26 coordinates; (3) the coordinates necessary for the notation are chosen, then angles are detected and (4) the skeleton is transcribed via the editing software

B. *Two camera approach for PCP improvement*

Next we illustrate the use of two cameras to increase the detection rate of the poses, as well as to recover more degrees of freedom for a rotation-invariant model of the human posture. This serves as an analogy to how motion descriptions are often explained by human instructors who teach movements, e.g. in martial arts, by using poses illustrated from two different angles. We observed that in a video from a single perspective some joints will almost inevitably be occluded by either a human body or an object, and therefore the software will take a (often incorrect) guess of the position of the limb, as can be seen in Figure 6.

We argue that by recording two videos at different angles an obstacle free view of many more frames can be provided, as can be seen in Figure 6. Therefore, the possibility of correct joint detection can be significantly increased by the two-camera approach since the two cameras have independent viewing angles. With a priori knowledge of the angles between the two cameras, it becomes possible to partially recover the correct posture even if the recognition fails in one camera. Furthermore, if more than one camera is used, a stereo-vision could be applied to create a rotation-invariant 3D model of the posture to create a coherent viewpoint of the motion sequence. For our experiment, two synchronized videos of the same motion were taken at two different angles. We labeled a frame as having detected the pose correctly, whenever at least one of the two cameras returned a valid estimation of the pose.

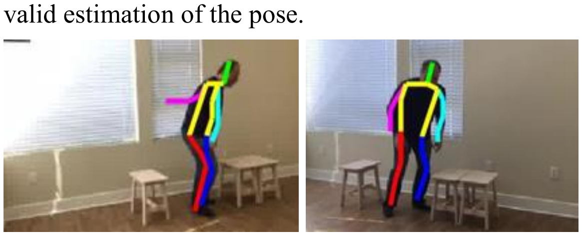

Figure 6. View from two cameras placed at an angle. The left arm is obstructed from the sight of the left camera, while the right camera has full view with correct pose estimation.

VI. RESULTS

A. *Joint coordinates with respect to time*

In this section we illustrate the nature of our data and explain the difficulty of selecting the representative motion primitives automatically. Figure 7 shows that the hand joints suffer from discrete jumps in the data, due to the inconsistency of correct detection leading to larger deviations within the data. Purely analytical approaches of graph analysis such as derivatives, without further signal processing, would lead to incorrect detections of appropriate key postures.

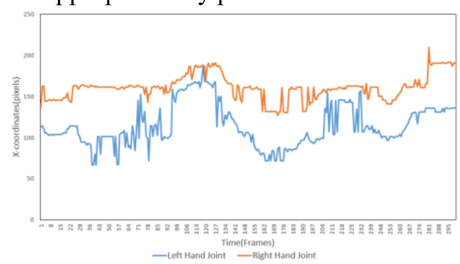

Figure 7. x-coordinates of the hand joints in respect to frames.

In order to automatically detect key postures from the human movement data, the signal must be further analyzed, e.g. by signal-processing techniques (low-pass filters etc.) However, we argue that as long as the key postures of motions can be correctly sampled, allowing a motion planner to interpolate between the samples to recover a whole motion, the selected frames suffice to serve the purpose of representing a movement for humans and robots to follow.

B. *Illustration of Motion Transcription*

Next we illustrate two successful cases of applying our method to automatically extract the characteristic frames from a video to transcribe the sequence to a HOM notation. Figure 8 shows video frames with the skeleton detected by the pose estimation software. Figure 9 shows the motion of Figure 8 transcribed in HOM notation. One can observe that the transcribed motion represents the recorded motion, and we argue that a humanoid robot would be able to interpret the script in order to reproduce the original movement. In Figure 10 we observe a situation in which body parts are occluded from camera view. In this case another camera could be used to provide additional information about the posture. We note in Figure 10 that while the first four frames have the correct detection of postures, the last four frames have errors in estimation due to parts of the limbs that are hidden from sight by the object or the torso. A second camera, viewing from a different angle, can be used to replace these frames, and recover the correct pose estimation. In Figure 11 we illustrate the synchronized frames from a different camera that can replace the frames and recover the correct estimation.

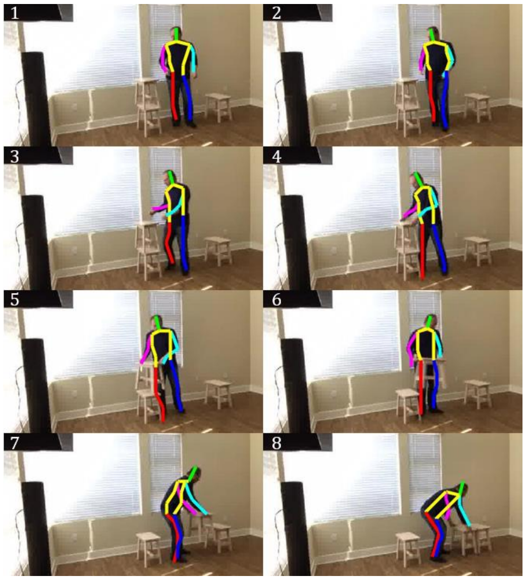

Figure 8. The detected skeletons from the sequence of images, where the motion is picking up a chair and putting it down in another location.

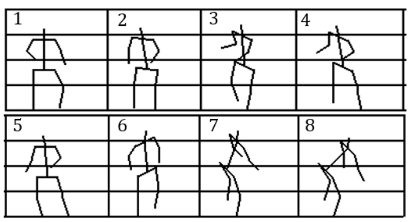

Figure 9. The program-transcribed HuRo script of the motion sequence illustrated by figure 8

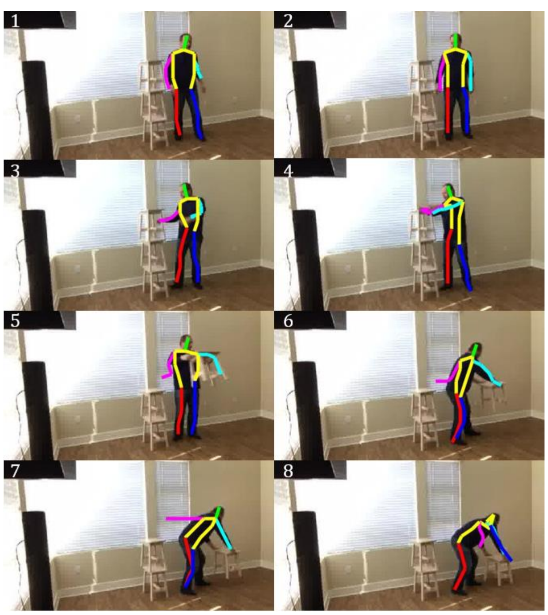

Figure 10. The detected skeletons from another motion of picking up a chair and putting it down in another location. In frames 5,6,7 and 8, the body parts are incorrectly detected due to poor line of sight.

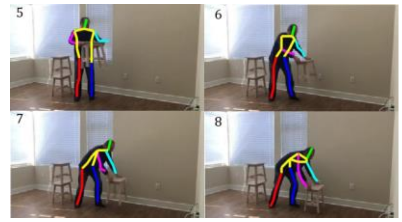

Figure 11. Alternative frames used for frames 5,6,7, and 8 that are taken from a different camera. These frames depict the poses at the same moments, but all the joints are detected correctly.

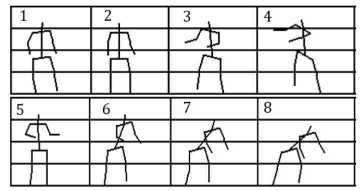

Figure 12. Transcribed HuRo script of the motion sequence illustrated by the first four frames of figure 12 (1,2,3,4), and the four frames of figure 13 (5,6,7,8). The top row and the bottom row use different cameras.

In order to connect the two sequences of postures into a coherent representation of a single motion (Figure 12), a method is required to rotate the skeleton sequence from the second camera to be compatible with the view from the first camera. This requires a rotation-invariant 3D model of the posture, which is not achievable in this case as we only have one valid image of the posture. However, under situations where we have multiple cameras filming at different angles, a coherent view of the sequences from a single angle can be recreated. Nonetheless, with enough processing even a single perspective may reach similar levels as a human observer.

C. PCP Analysis of using two cameras

Without an annotated ground-truth of the joints, we utilized a manual PCP scoring method where each instance of detection was compared to recognition by human perception. We expect a general decrease in accuracy for our videos, as many of the frames may include poses where a part of the limb may be hidden from sight by another body part or an obstacle. Individual scores for 100% detection (when all the joints are correctly detected) measured from two videos recorded by two cameras from different angles are represented in Table 1 and Table 2 (V for video, C for camera)

Dataset	100% Correct	Total	Score
V1C1	84	298	**28.19**
V1C2	83	298	**27.85**
V1 Combined	140	298	**46.98**

Table 1. The rates for 100% detection of joints for two cameras in video 1.

Dataset	100% Correct	Total	Score
V2C1	112	350	**32.00**
V2C2	68	350	**19.43**
V2 Combined	161	350	**46.00**

Table 2. The rates for 100% detection of joints for two cameras in video 2

From Table 1 and Table 2, we can observe that when analyzing for 100% correct detection, the scores improve by a factor of 1.7 on average, which means that the correct pose was detected 1.7 times more often when using two cameras instead of one. Throughout the analysis we noted that many of the frames were wrong by one of the arms, since arms were most likely to be obstructed by the torso. Chen and Yuille [15] compare in their most recent work PCP scores for different body parts, and it is observed that the arms have much lower PCP scores compared to other parts. As such we also provide the PCP scores that tolerate an error rate of 20%, specifically choosing the frames that are incorrect by a single arm.

Dataset	80% Correct	Total	Score
V1C1	140	298	**46.98**
V1C2	152	298	**51.01**
V1 Combined	217	298	**73.82**

Table 3. The rates for 80% detection of joints for two cameras in video 2, specifically excluding one of the arms.

Dataset	80% Correct	Total	Score
V2C1	198	350	**56.57**
V2C2	119	350	**34.00**
V2 Combined	250	350	**71.43**

Table 3. The rates for 80% detection of joints for two cameras in video 2, specifically excluding one of the arms.

Although complete detection of two arms is often critical for correct execution of a desired movement, incomplete detection can still be meaningful as the data provides information about intermediate joints that may be used to interpolate and recreate a whole motion from scripts. One should note that the use of multiple cameras in posture estimation would benefit automatic extraction of representations in any notations, not only in HOM.

VII. CONCLUSIONS

The paper introduced a novel movement writing system oriented to humanoids. It presented a manual editorial tool for specifying a new script, and the first automated method of transcribing human motion from a video to a movement writing notation. Video is preferable to Kinect or motion capture systems as it can be used in outdoors, at a distance and non-intrusive to humans. The paper also illustrates that through the use of two cameras, we can improve the PCP rate by a factor of 1.7, enhancing the automation rate at which the motion is correctly transcribed into a movement writing script. We discuss the current limitations of our system in the process of automation and detection; further improvement is possible by utilizing even more cameras, or improved estimation techniques to obtain more robust and rotation-invariant sequences of motion. The automated detection of the characteristic postures is also a possible improvement, which may be done through the filtering and analysis of the joint positions.

HOM aims to be an easier and more intuitive movement writing system then other proposed systems. As the notation gets extended it is possible to include objects and interactions into our scripts. State-of-art object recognition systems such as Faster RCNN [16] or simple edge-detection algorithms could be used to detect an object from a video and transcribe them to motion scripts, specifying the object of interest for humans and robots. Additionally, the manner of interaction (i.e. speed, power, etc.) with the object could be specified through symbols and verbal descriptions that follow the scripts. Finally, HOM provides both an intuitive interpretation of motion, as well as the basis from which a holistic motion could be interpolated. As such, it is particularly useful in expressing multi-agent situations with cooperative activities between humans and robots. HOM can be extended to describe these situations by incorporating multiple actions and objects into the scripts.

ACKNOWLEDGMENTS

This research was performed at the Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA 91107, under a contract with the National Aeronautics and Space Administration (NASA). The authors wish to thank the anonymous reviewers for their helpful suggestions to the early manuscript.

REFERENCES

[1] Valerie Sutton, “Sutton Movement Shorthand, Book One: The Classical Bellet Key”, Sutton Movement Writing Press, Nob 1 1973
[2] Y. Yang, D. Ramanan. ”Articulated Pose Estimation using Flexible Mixtures of Parts” Computer Vision and Pattern Recognition (CVPR) Colorado Springs, Colorado, June 2011.
[3] Von Laban et al, “Laban’s Practice of Dance and Movement Notation”, Princeton Book Co Pub, 1975
[4] Benesh et al, “Reading Dance: The Birth of Choreology”, International Specialized Book Services, 1997
[5] T. Nakata et al, “Expression of Emotion and Intention by Robot Body Movement”, Proceedings of International Autonomous Systems 5, 1998
[6] J. T. Butler and A. Agah, “Psychological Effects of Behavior Patterns of a Mobile Personal Robot”, Autonomous Robots 10, 2001, pp.185
[7] H. Knight and R. Simmons, “Laban Head-Motions Convey Robot State: A Call for Robot Body Language”, IEEE Conference on Robotics and Automation (ICRA), 2016
[8] A. Samadani et al, “Laban Effort and Shape Analysis of Affective Hand and Arm Movements”, Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), 2013
[9] K. Hachimura et al, “Analysis and Evaluation of Dancing Movement Based on LMA”, IEEE International Workshop on Robots and Human Interactive Communication, 2005
[10] T. Lourens et al, “Communicating Emotions and Mental States to Robots in a Real Time Parallel Framework using Laban Movement Analysis”, Robotics and Autonomous Systems, 2010, pp.1256-1265
[11] S.S. Kim et al, “Whole-body Motion Imitation using Human Modeling”, Proceedings of IEEE International Conference on Robotics and Biomimetics, 2008
[12] B. Dariush et al, “Whole Body Humanoid Control From Human Motion Descriptors”, IEEE International Conference on Robotics and Automation (ICRA), 2008
[13] P. Felzenszwalb and D. Huttenlocher. “Pictorial structures for object recognition” International Journal of Computer Vision, 61(1):5579, 2005
[14] D. Ramanan. “Learning to parse images of articulated bodies”,Advances in Neural Information Processing Systems (NIPS), 19:1129, 2007
[15] X.Chang and A.Yuille, ”Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations”, Advances in Neural Information Processing Systems (NIPS), 2014
[16] S. Ren et al, “Faster R-CNN: Towards real- time object detection with region proposal networks”, Advances in Neural Information Processing Systems (NIPS), 2015.