Towards Human Pose Semantic Synthesis in 3D based on Query Keywords

Mo'taz Al-Hami and Rolf Lakaemper
Department of Computer & Information Sciences, Temple University, Philadelphia, PA 19122, U.S.A.

Keywords: Human-pose, 2D Human-pose Estimation, 3D Human-pose Reconstruction, Silhouette Value, Hierarchical Clustering.

Abstract: The work presented in this paper is part of a project to enable humanoid robots to build a semantic understanding of their environment adopting unsupervised self-learning techniques. Here, we propose an approach to learn 3-dimensional human-pose conformations, i.e. structural arrangements of a (simplified) human skeleton model, given only a minimal verbal description of a human posture (e.g. "sitting", "standing", "tree pose"). The only tools given to the robot are knowledge about the skeleton model, as well as a connection to the labeled images database "google images". Hence the main contribution of this work is to filter relevant results from an images database, given a human-pose specific query words, and to transform the information in these (2D) images into a 3D pose that is the most likely to fit the human understanding of the keywords. Steps to achieve this goal integrate available 2D human-pose estimators using still images, clustering techniques to extract representative 2D human skeleton poses, and the 3D-pose from 2D-pose estimation. We evaluate the approach using different query keywords representing different postures.

1 INTRODUCTION

Humanoid robots are increasingly adapting to physically mimic human actions and poses, which require an understanding of human poses and how they can be captured with relation to the robot's joint movements. While the translation of a known target pose to (rotational) joint positioning is a generally solved problem in inverse kinematics, this paper aims to generate a target pose description, given a simple verbal description ("stand", "warrior pose"). It therefore aims at adding semantic content to a given query words and their relation to the robot's skeleton model, which represents the robot's physical self-understanding. In short, this approach enables the robot to experience the posture related meaning of words descriptions. Other than supervised learning settings, e.g. simulations of children's training and learning (Ikemoto et al., 2012), this project is based on unsupervised self-learning. While the (Ikemoto et al., 2012) project focuses on improving the interaction with human counterpart, our project is more pragmatic in the sense that it should quickly and efficiently enable the robot to understand its environment. With such a goal in mind, extending the humanoid robot's capability to support self-learning is, for obvious reasons (no supervision needed, no ground truth labeled data needed), an important step.

Recent works (Mu and Yin, 2010), (Jokinen and Wilcock, 2014) started to improve Internet-based spoken dialogue between a human and a humanoid robot. The humanoid robot's side in this conversation uses an active source, Wikipedia, to improve spoken dialogue capabilities. This kind of research opens the door to think about a different data source, e.g. images, to act as the basic descriptor for posture relating tasks. The motivation for using images is their ability to provide a visual description of a query's keywords. This style of analyzing queries could also handle conceptual issues for many applications like (Mu and Yin, 2010), (Jokinen and Wilcock, 2014).

For non-rigid objects like human-pose, it is important to understand the relation between the search query (i.e. keywords representing a human-pose conformation) and its related pose conformation, such that there is a clear description for this query keywords using images. For example, we want the robot to learn the posture "standing pose" without any human intervention. The robot is only allowed to query a general 2D images database, here google images. From the image results that google returns from the query "standing pose", the robot needs to decide, what

Al-Hami M. and Lakaemper R..
Towards Human Pose Semantic Synthesis in 3D based on Query Keywords.
DOI: 10.5220/0005258704200427
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 420-427
ISBN: 978-989-758-091-8

the skeleton expression of "standing pose" is. Applying the search on a "standing pose" query would retrieve a group of images which are strongly related to the query "standing pose", a group of images which are weakly related to the query, and a group of outliers (false positive images) which are completely unrelated. The question now is how to capture the query related images from the set of all retrieved images. Next, how to establish a single general pose from these images which is able to capture the main properties of the "standing pose" keywords.

In this paper, we propose a framework which uses query keywords related to specific human-pose conformation as our input and returns back a 3D human-pose as output. The importance of this work is its ability to bridge the gap between estimating the human pose in 2D images and constructing 3D human pose from 2D image. To clarify and summarize the approach:

After using a query keywords as an input to google, and working on the retrieved images, the proposed framework focuses on processing all the retrieved images and extract a reliable poses from these images. Second, based on the extracted poses, approximate one general pose as a representative pose for the query. Finally, construct a 3D human pose related to the approximated 2D pose. This kind of self awareness allows humanoid-robots to perform a self-motivated tasks rather than predefined ones.

Our main contributions in this work can be summarized as follows: (1) Propose a hierarchical binary clustering approach intended to filter poses iteratively in order to get a consistent and representative cluster of poses. (2) Approximate the representative cluster of poses using one general pose summarizing the query keywords semantic. The paper is organized as follows. Related work is described in Section 2. Human-pose estimation process is discussed in Section 3. Extracted human-poses clustering and evaluation are discussed in Section 4 and 5. Finally, conclusion remarks appear in Section 6.

2 RELATED WORK

Human-pose consists of a set of related parts, and these parts are arranged according to a valid hierarchical structure. Human-pose estimation in still images focuses on how to detect a pose and localize its parts within a 2D image. Many works like (Ramanan and Sminchisescu, 2006), (Ramanan, 2006), (Fergus et al., 2003), (Ioffe and Forsyth, 2001), (Ferrari et al., 2008) use deformable model to create part based templates, and discover the relationship between parts. The power of this approach comes from its ability to detect pose without a prior knowledge about the background or appearance (i.e. clothes, skin).

Shape matching has been applied in (Gavrila, 2000) to generate a shape template using a distance transform approach. This template is intended to capture the objective shape variability. In (Ren et al., 2005) pairwise constraints between human-pose parts supported with an image segmentation approach is used. To locate the candidate parts in a 2D image, a bottom-up approach is applied.

In (Mori and Malik, 2002), the work provides an attempt to construct a 3D human-pose model. In addition to an unseen 2D image, the approach uses a set of labeled examplers representing different viewpoints with respect to the used camera. Next, the approach finds a sufficient match for the unseen image in the examplers set, and then a 3D model is constructed based on the labels of these sufficient examplers. The work in (Yao and Fei-Fei, 2010) focuses on studying activities that include human object interaction like many sport activities (i.e. tennis, football). The work proposes a mutual context approach between human-pose and objects. This mutual context facilitates the recognition process of an object and the estimation of a human-pose as well.

Mixed Bayesian model has been used in (Lan and Huttenlocher, 2004). The approach consists of hidden Markov model and pictorial structure to estimate human-pose. Since human-pose has a high symmetry between limbs (i.e. right arm with left arm, and right leg with left leg), work in (Lan and Huttenlocher, 2005) incorporates this symmetry with tree structure by taking into account balancing symmetric limbs. In (Eichner et al., 2012), the work focuses on detecting upper body parts and uses the deformable model with extension to reduce the search space for parts localization. The work uses some prior assumption about the human-pose. These assumptions describe the relation between the head and the torso and their upright appearance.

For more human and robot interaction, work in (Jokinen and Wilcock, 2014) improves social human robot communication through using a WikiTalk application. This application provides a social conversation system aided with Internet based capability (i.e. connected to Wikipedia). The humanoid robot NAO was used as a testing platform for the WikiTalk and improved with gesturing movements. The work in (Al-Hami and Lakaemper, 2014) applies the genetic algorithm on NAO humanoid robot pose to fit well on unknown sittable objects including (boxes and balls).

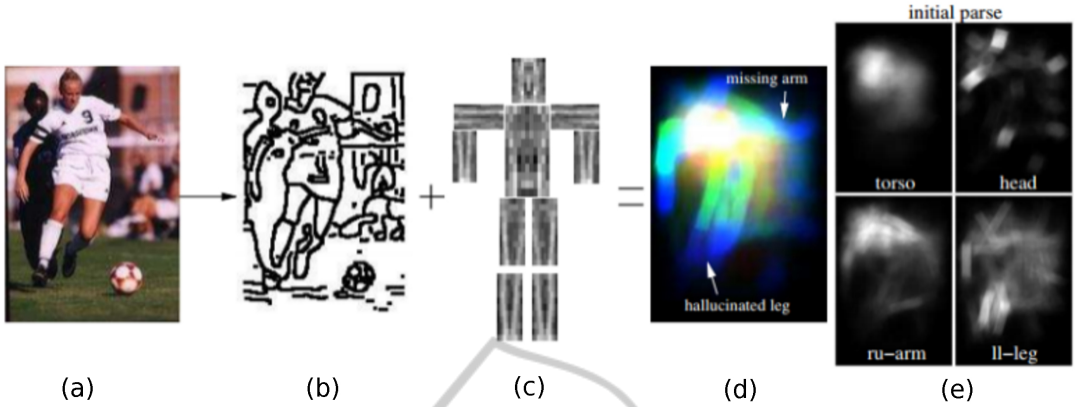


Figure 1: Deva Ramanan pictorial structure deformable model approach for human-pose estimation (Ramanan, 2006). (a) original image, (b) constructed edge map from the original image, (c) Spatial priors (templates) are used based on the learned spatial priors from a database of labeled joints images, (d) initial parsed model for the whole pose, and (e) initial parsed model for independent joints.

(a) Standing (b) Sitting (c) Standing (d) warrior

Figure 2: Some examples of human-pose estimation results when applying approach in (Ferrari et al., 2008) on retrieved images by google for different pose conformations. The labels under figures refer to the used query keywords for these images.

3 HUMAN-POSE ESTIMATION

Human-pose consists of a non-rigid articulated object having a large number of degrees of freedom. Such pose structure (shown in Table 1) leads to a high dimensional space of possible pose conformations. Pictorial structure introduced by Martin Fischler in (Fischler and Elschlager, 1973) has been used to represent human-pose parts model. The approach takes into account the spatial conformation and the parts appearance of a pose within an image. For the sake of clarity, assume that the human-pose L consists of m joints $L = \{l_1, l_2, ..., l_m\}$ where l_i is the location of i^{th} joint in that pose. Each joint in the model has a specific conformation describes its appearance in a 2D image and identified by the conformation triple $[x_i, y_i, \theta_i]$. (x_i, y_i) represents the location and θ_i is the orientation of that joint. The goal of such a model is to detect a human-pose/poses in an unseen 2D image, then to find each part in each pose appears in that image.

3.1 Spatial Priors

The process of recognizing a human-pose in a 2D image requires a knowledge about each joint exists in this pose conformation. Machine learning approaches have been applied on a set of labeled images including human-poses. This set of images is labeled human-pose parts. The task is to learn the localization of each part in a pose such that the localization should satisfy some criteria like maximum likelihood (ML) solution (Ramanan, 2006). The output of this learning process is a general spatial prior template (local information template describes part appearance in an image and its localization) for the human-pose like the one shown in Fig. 1(c).

In the spatial priors model, a human-pose uses parameters which capture two main aspects in the model, spatial relationship S, and appearance A. For such model $M = (S, A)$, the required parameters are learned using machine learning approaches using a data set of labeled images. The process of building spatial priors (i.e. template for each joint in the human-pose which favors a certain triple $[x_i, y_i, \theta_i]$) is concerned with how to determine the $P(L)$ as a prior distribution.

3.2 Iterative Parsing Approach

Iterative parsing approach proposed by Deva Ramanan (Ramanan, 2006) is a general approach (i.e. does not depend on specific features like clothes, color, or skin) which is able to recognize a human-pose's joints in a 2D image. The contribution in this approach is its ability to detect and localize a pose's joints using iterative parsing process applied on a 2D image. Throughout each iteration a better features can

be learned and tuned to a specific 2D image, resulting in an improved pose detection and joints localization.

In the first parse iteration, a spatial priors are used to produce soft (i.e. roughly) estimate for a pose joints. This estimate is rough since the spatial priors are trained on a general labeled database of 2D images. After the first parsing iteration, the spatial priors starts to be tuned towards the current image features (i.e. the iterative parsing approach tune the general template to more specific one related to the current image).

The approach in Fig. 1 focuses on building joint-based region model which determines each human-pose joint location. In the first iteration, edge map is used with the spatial priors to generate the initial soft parse, after that the regions generated in each parse iteration are used to enhance the pose estimation process in the next iteration.

For estimating human-pose joints, the approach emphasizes the importance of the conditional likelihood model, rather than the classical maximum likelihood model. For conditional likelihood the goal is:

$$\max\ P(L\backslash I,\Theta) \qquad (1)$$

which means maximizing the likelihood estimate that produces the best estimate when applied to labeling a human-pose parts in a given 2D image I. Based on a group of labeled training images, the learned model Θ focuses on learning the model parameters which satisfy this conditional likelihood model. Assuming the relation between a pose's joints can be represented as a tree structure where there is a relation between each joint and its parent. Using such structure, the log-linear deformable model can be represented by the conditional likelihood model:

$$P(L\backslash I,\Theta) \propto\ exp(\sum_{i,j\in E}\psi(l_i,l_j)+\sum_{i}\phi(l_i)) \qquad (2)$$

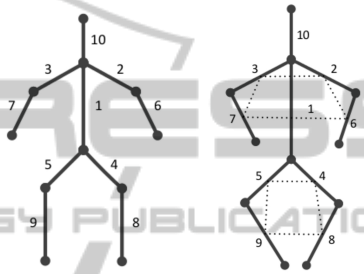
In this model, $\psi(l_i,l_j)$ shows the spatial prior for joint l_i using its relative arrangement to joint l_j which both have a connection in the tree structure. This prior captures the joint spatial arrangement. The term $\phi(l_i)$ captures appearance prior for each joint, and serves as a local image evidence for these joints. For more details about this log-linear deformable model see (Ferrari et al., 2008), (Ramanan and Sminchisescu, 2006).

3.3 Poses Skeletons Extraction

Applying the iterative parsing approach as mentioned earlier on the retrieved images would produce a database of extracted poses skeletons. Many implausible skeletons are excluded from the database

Table 1: The used tree structure for representing a human-pose.

Number	Joint	Parent
01	Torso	-
02	Left Upper Arm	Torso
03	Right Upper Arm	Torso
04	Left Upper Leg	Torso
05	Right Upper Leg	Torso
06	Left Lower Arm	Left Upper Arm
07	Right Lower Arm	Right Upper Arm
08	Left Lower Leg	Left Upper Leg
09	Right Lower Leg	Right Upper Leg
10	Head	Torso



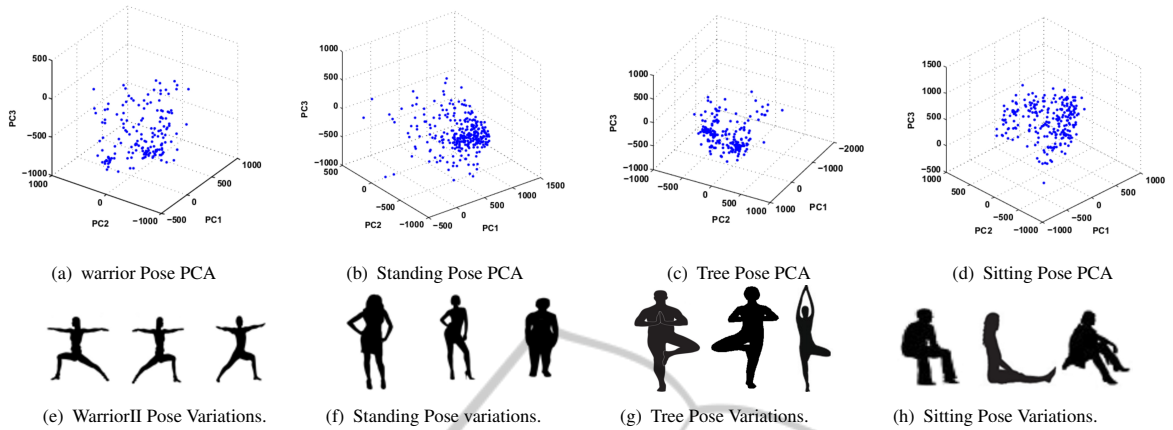
(a) Pose Structure (b) Feature Space

Figure 3: **(a)**: Human-pose structure used in this work. **(b)**: The used distances which appended in the feature vector.

(we kept only poses where the distances between a pose joints according to the tree structure is less than distance σ threshold). Applying the pose estimator on the retrieved images revealed many limitations. Some examples from the retrieved results are shown in Fig. 2. We noticed some wrong segmentation (separating the human-pose from the background in a 2D image) for many images (see 2(b)). Another limitation happens in a human-pose estimation even with correct segmentation, where the estimator localizes wrongly some of a pose joints (see 2(c)). However, the estimator was able to segment and estimate many poses correctly like ones shown in 2(a), 2(d). Since human-poses in the retrieved images appear at different depths and different scales, all extracted human-poses are scaled to a one unified bounding box.

4 EXTRACTED HUMAN-POSES CLUSTERING

The poses extraction process suffers from limitations mentioned earlier. To address such problems, a hierarchical binary clustering approach is used to extract a sufficient representative set of human-poses skele-



(a) warrior Pose PCA (b) Standing Pose PCA (c) Tree Pose PCA (d) Sitting Pose PCA

(e) WarriorII Pose Variations. (f) Standing Pose variations. (g) Tree Pose Variations. (h) Sitting Pose Variations.

Figure 4: Poses appearance in space after scaling to a predefined bounding box. (a)-(d) show the plot of the first three principal components of the poses after transforming the poses space into a new space using PCA analysis. (e)-(h) show a common typical variation for each pose in the real life.

tons. The basic idea depends on the existing consistency between poses skeletons which are strongly related to the query keywords. Using a consistency as a criteria for the selected poses skeletons, we can focus our attention towards finding a consistent representative set from human-poses skeletons. Since the available database (i.e. retrieved human-pose skeletons) are high dimensional, the consistency is coupled with poses spatial localization in the space. This spatial localization in the space depends on the used feature vector structure (described later).

In Fig. 4 we used principal component analysis (PCA) to show the database of human-poses in the space for different pose conformations as well as their typical variation in real life. By plotting the first three principal components of the used feature vector, the localization of poses shows some areas where there are a high density spots of poses while others are spread out randomly.

Highlighting the cluster consistency between the cluster members can offer clues about the main characteristics of a pose conformation. By using a suitable feature vector (described later) and using the iterative hierarchical binary clustering approach using kmeans clustering agorithm, we can cluster the poses skeletons. The clustering approach separates the poses points iteratively into two clusters, one of them (which has a higher consistency value) is selected to be the cluster of interest while the other one is discarded. The clustering procedure is repeated multiple times on the selected cluster until the cluster consistency value (Silhouette value) is more than a threshold t (see Fig. 5). Obtaining a consistent representative cluster provides a subset of related real poses.

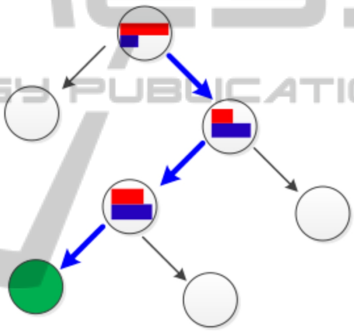


Figure 5: Hierarchical binary clustering approach. At each level, poses skeletons are divided into two clusters, one of them which is more consistent is selected while the other one is discarded. The process is repeated until the representative cluster consistency exceeds a threshold value t. Kmeans clustering algorithm is used in this approach.

4.1 Feature Vector

Kmeans clustering approach separates data points (poses) based on their proximity to clusters centroids. The used feature vector structure for representing a pose is described in Table 2. The feature vector consists of two main parts. The first part keeps center points of a pose joints as shown in Fig. 3(a) while the second part keeps distances between a pose joints as shown in Fig. 3(b). The motivation of using such structure can be described as follows. The first part captures a pose joints absolute spatial location using center points, while the second part captures a pose joints locations relative to their parent joints and their peer joints as well.

Table 2: Structure of the used feature vector consists of two parts, one of them captures the spatial localization (center points) while the other one captures the joints relations (distances). Numbers in the first part refer to center points of specified joints (i.e. x offset and y offset of a center point). d[a,b] in the second part refers to the Euclidian distance between center points of joints a and b.

Feature vector structure																
Absolute spatial arrangement										Relative spatial arrangement						
Center points										Distances						
01	02	03	04	05	06	07	08	09	10	$d[2,3]$	$d[6,7]$	$d[2,6]$	$d[3,7]$	$d[4,5]$	$d[8,9]$	$d[4,8]$	$d[5,9]$

4.2 Silhouette Value

Silhouette value analysis measures a cluster quality and consistency. In this analysis each pose in a cluster is assigned a normalized similarity value (S_i). This value reflects a pose similarity to poses in the same cluster when compared to poses in the other cluster.

$$S_i = (b_i - a_i) / \max(a_i, b_i) \quad (3)$$

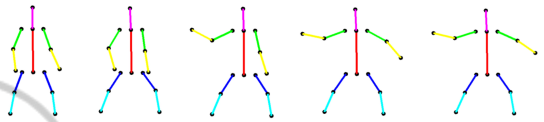
The similarity is captured through average Euclidian distance from a given pose to all other poses in the same cluster (a_i). The other criteria which is the average Euclidian distance from a pose to all poses from the other cluster (b_i). Silhouette values range is between [-1,1]. To assign a cluster a single silhouette value, we use the average of all poses silhouette values in that cluster as representative silhouette value. The resulted value is used as the criteria in the hierarchical binary clustering. It decides which cluster is the relevant one and which cluster has to be discarded.

4.3 Approximate Pose Skeleton Model

Model approximation captures the general characteristics of a cluster. By approximation, each joint in the approximate pose skeleton represents the median orientation of that joint of all poses within a cluster independently of other joints. Median is chosen since average is sensitive to outliers. The approximation process enables us to track the cluster changes at each level in the hierarchical binary clustering.

4.4 3D Human-pose Reconstruction

The work in (Ramakrishna et al., 2012) constructs a 3D human-pose using a 2D image with landmarks located at specific positions and specifying positions of a human pose in that image. The approach uses a projected matching pursuit algorithm in order to minimize the error amount between the original landmarks in the 2D image and the projected ones from the constructed 3D model. Assuming a fixed intrinsic camera parameters, the approach is able to approximate the extrinsic camera parameters and a 3D model using a corpus (database) of 3D poses. In this work, we use approximate poses generated from representative clusters to be used for 3D reconstruction. Since the joints positions (x offset and y offset) are already known, we can automatically mark the required landmarks.



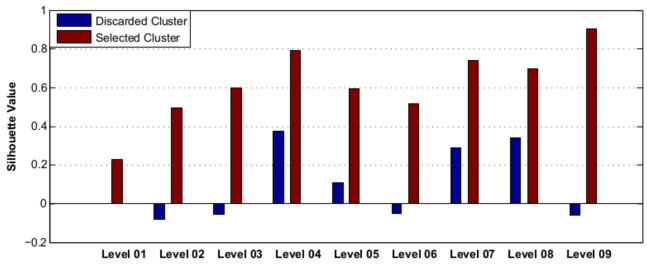
(a) L 01 (b) L 02 (c) L 03 (d) L 04 (e) L 05

Figure 6: Warrior approximation pose at different hierarchical binary clustering levels.

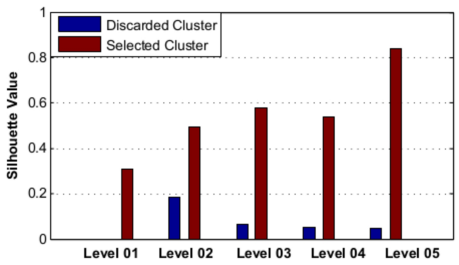
5 EVALUATION

In experiments, we try to show the effect of the used hierarchical binary clustering approach in producing a representative subset of poses. The approach has been applied to three different queries. Since the extracted human-poses from the retrieved images (varies between 700 to 900 images in our experiments) might include implausible poses, we filtered the extracted human-poses to plausible ones (the Euclidean distance between related joints in the hierarchy is below that a threshold value). Using the filtered set of poses the hierarchical binary clustering approach is applied iteratively. The goal is to cluster poses until we get a consistent cluster (average cluster silhouette value is at least 0.8). The first query was "warrior pose". After applying the hierarchical binary clustering approach 5 levels, we were able to get a consistent cluster. Fig. 6 shows the approximate pose at each level in the clustering approach. As far as we perform more iterations in the clustering approach, the pose conformation moves towards a sufficient representative approximate pose for the original query keywords "warrior pose".

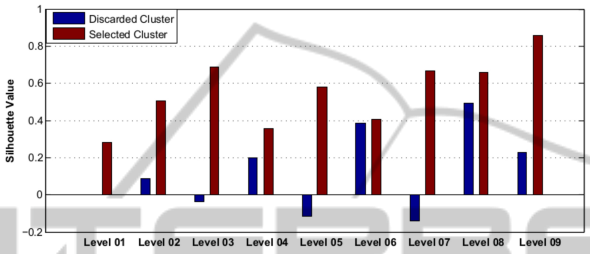
The second query in the experiment was "tree pose". Following the same approach, after 9 levels of clustering we were able to reach a consistent cluster. Fig. 7 shows the approximate pose at each level in the clustering approach. The approximate pose is able to



(a) Tree pose hierarchical binary clustering Silhouette values

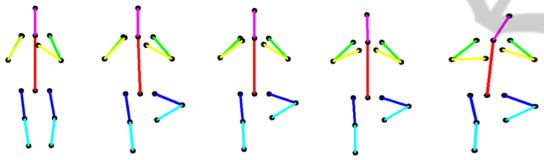


(b) warrior pose hierarchical binary clustering Silhouette values



(c) Standing pose hierarchical binary clustering Silhouette values

Figure 10: Average Silhouette value at each level in the hierarchical binary clustering approach for the selected and discarded clusters for three different queries. The stop condition for the clustering process is an average Silhouette value exceeds 0.8.

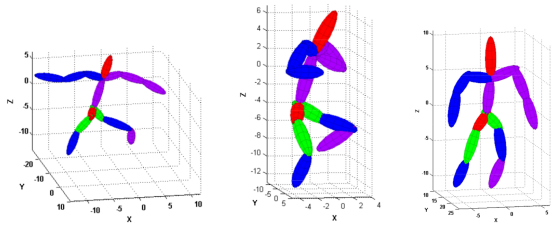


(a) L 01 (b) L 03 (c) L 05 (d) L 07 (e) L 09

Figure 7: Tree pose approximation pose at different hierarchical binary clustering levels.

(a) L 01 (b) L 03 (c) L 04 (d) L 07 (e) L 09

Figure 8: Standing pose approximation pose at different hierarchical binary clustering levels.



(a) Warrior pose (b) Tree pose (c) Standing pose

Figure 9: Reconstructing 3D human-pose models based on approximate 2D human-pose models.

capture the common characteristic of a real tree pose. The third query was "standing pose". After 9 levels of clustering we got a consistent cluster as shown in Fig. 8.

In Fig. 10, average silhouette values for the selected and discarded clusters are measured at each clustering level for queries used in each experiment. As the silhouette value decreases, the amount of inconsistent poses appearing within a cluster increases.

The last stage uses obtained results from the hierarchical binary clustering approach and translates them into an approximate 3D human-pose model. The motivation is to build a 3D model which allows us to get an automatic visual translation of a query keywords of a human-pose conformation. Using the approach in (Ramakrishna et al., 2012), we applied the projected matching pursuit algorithm on approximate poses. Fig. 9 shows the obtained 3D human-pose model for each approximate pose.

A case where this approach does not work well, is when the iterative parsing approach mentioned in section 3 produces a weak and inaccurate estimation. In the case of sitting pose, the joints might have many overlaps, and this makes the estimation process weak in a majority of cases. A clear example of the estimated poses are shown in 2(b). In addition to that, the variations that might occur in the sitting pose is also high (person might sit on the floor, on different types of chairs, or might sit with cross legged). Applying the hierarchical binary clustering approach on

the estimated poses in this case would not produce a sufficient representative cluster.

6 CONCLUSION

In this paper, we presented a hierarchical binary clustering approach which bridges the gap between 2D human-pose estimation in an image, and 3D human-pose reconstruction from 2D human-pose skeleton. Using this approach we are able to translate query keywords representing a human-pose conformation into an approximate 3D human-pose skeleton. The work in (Al-Hami and Lakaemper, 2014) uses the genetic algorithm to adjust a humanoid robot pose, so the robot would fit well on an unknown sittable object. Extending such approach could be accomplished by allowing the humanoid robot to adopt self-leaning using simple verbals describing a specific human-pose. This style of query analysis allows us to extend the humanoid robots ability toward self-motivated learning, allowing them to move forward in many applications. We discussed a hierarchical binary clustering approach to extract a consistent representative subset of human-poses. Silhouette value measurement was used to capture a cluster consistency, and cost transformation was used to rank poses within a cluster according to their closeness to the approximate model. For future work, we want to improve pose estimation accuracy in 2D images. Also we want to improve the 3D reconstruction performance by forcing joint valid rotation ranges, such that the constructed 3D pose is within a valid joints rotations.

REFERENCES

Al-Hami, M. and Lakaemper, R. (2014). Sitting pose generation using genetic algorithm for nao humanoid robot. In *IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), 2014*, pages 137–142. IEEE.

Eichner, M., Marin-Jimenez, M., Zisserman, A., and Ferrari, V. (2012). 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, 99(2):190–214.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2003*, volume 2, pages II–264. IEEE.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008*, pages 1–8. IEEE.

Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92.

Gavrila, D. (2000). Pedestrian detection from a moving vehicle. In *Computer Vision ECCV 2000*, pages 37–49. Springer.

Ikemoto, S., Amor, H. B., Minato, T., Jung, B., and Ishiguro, H. (2012). Physical human-robot interaction: Mutual learning and adaptation. *Robotics & Automation Magazine, IEEE*, 19(4):24–35.

Ioffe, S. and Forsyth, D. A. (2001). Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68.

Jokinen, K. and Wilcock, G. (2014). Multimodal open-domain conversations with the nao robot. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 213–224. Springer.

Lan, X. and Huttenlocher, D. P. (2004). A unified spatio-temporal articulated model for tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2004*, volume 1, pages I–722. IEEE.

Lan, X. and Huttenlocher, D. P. (2005). Beyond trees: Common-factor models for 2d human pose recovery. In *Tenth IEEE International Conference on Computer Vision (ICCV), 2005*, volume 1, pages 470–477. IEEE.

Mori, G. and Malik, J. (2002). Estimating human body configurations using shape context matching. In *Computer Vision ECCV 2002*, pages 666–680. Springer.

Mu, Y. and Yin, Y. (2010). Human-humanoid robot interaction system based on spoken dialogue and vision. In *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), 2010*, volume 6, pages 328–332. IEEE.

Ramakrishna, V., Kanade, T., and Sheikh, Y. (2012). Reconstructing 3d human pose from 2d image landmarks. In *Computer Vision ECCV 2012*, pages 573–586. Springer.

Ramanan, D. (2006). Learning to parse images of articulated bodies. In *Advances in Neural Information Processing Systems*, pages 1129–1136.

Ramanan, D. and Sminchisescu, C. (2006). Training deformable models for localization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 1, pages 206–213. IEEE.

Ren, X., Berg, A. C., and Malik, J. (2005). Recovering human body configurations using pairwise constraints between parts. In *Tenth IEEE International Conference on Computer Vision (ICCV), 2005*, volume 1, pages 824–831. IEEE.

Yao, B. and Fei-Fei, L. (2010). Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010*, pages 17–24. IEEE.