Pose Description Based on Natural Relation Sets

Benjamin Raynal, Vincent Nozick

Pose Description Based on Natural Relation Sets

Benjamin Raynal
A3SI LIGM
Université Paris-Est
Email: raynal@univ-mlv.fr

Vincent Nozick
A3SI LIGM
Université Paris-Est
Email: vnozick@univ-mlv.fr

***Abstract*—Motion capture and pose estimation of human beings is a highly active research area and is related with various applications in many different fields such as 3D character animation, surveillance or human-machine interfaces. A specific problem of these two last applications is the pose and motion estimation of the subject, i.e. understanding which action, among a predefined set, the subject is performing. In this paper, we propose a new high level pose description method based on set of relative body part information, easily understandable by humans, such "above/below" or "between/on each side of". From this pose description, we introduce two kinds of usage: the recognition of a pose described by the user and the detection of poses similar to a set of samples.**

I. Introduction

Human motion analysis is a very active and still challenging research topic that involves many applications ranging from computer interfaces to human behavior understanding. Both acquisition device and analysis method depend on the application and the level of abstraction required. Some applications such as virtual reality or sign language require fast and accurate elementary motion description whereas human behavior description and understanding or content-based video retrieval may involve more complex gesture analysis to recognize human motion patterns. Human motion analysis is also highly related to visual surveillance techniques that is also a very active research topic.

Motion analysis is strongly related to motion capture, that can be performed from specific devices as well as from a single 2D video sequence. Moeslund et al. [1] present a very complete overview of recent motion capture techniques specifying if the methods estimate the pose and perform recognition. For more information about motion capture, this article also refers to previous surveys in the field.

Among all the gesture analysis methods overviews, we can highlight the survey of Mitra and Acharya [2] that presents recent works on the analysis of hand and arm gesture as well as on facial expression. Ji and Liu [3] also present an overview of view-invariant human motion analysis, dealing with the methods that remain unaffected by different view points of the camera. They distinguish the pose representation and estimation (i.e. how to estimate a 3D pose from individual image in a sequence) to the action representation and recognition (i.e. how to estimate a human action pattern).

More generally, we can differentiate the cases where the input data is computed from a 2D video stream, from a multi-view method or from a specific motion capture device. We can also notice that most of the gesture recognition methods are based on markov chains, neural networks, particle filtering or statistical modeling.

As a relevant paper in gesture analysis, we can cite Agarwal and Triggs [4] who propose a learning based method from a sequence of images using a nonlinear regression on the user shape variations. Ryoo and Aggarwal [5] introduce a hierarchical spatio-temporal relationship matching that overcomes restrictions on periodic actions. Shakhnarovich et al. [6] mix an efficient hashing function with a learning based pose estimation method from a large database. Sullivan and Carlsson [7] propose a method to match a 3D geometric data with a specific human action, even from single frame postures.

In this article, we present a new pose description approach using set theory as a mathematical support. The pose description is based on a set of local relationships between human body parts. Put together, these local relationships design a global description of the subject. Furthermore, these relationships are designed to be directly understandable by humans expressed in natural language. These characteristics provide an interesting framework for the description of a specific pose in natural language or by a set of samples.

The rest of the paper is organized as follows: in Sec. II, we define what information we have in input and explain the choice of the features we use for our descriptor. In Sec. III and Sec. IV, we introduce respectively the axis-based and betweenness relations. Once our descriptor is fully described, we propose different similarity measurement on it in Sec. V, and a syntax for its conversion in natural language in Sec. VI. Finally, we provide some results of their application in Sec. VII.

In the following parts, we refer to "human body parts" any elements to be identified and described by a 3D position and a name. Indeed, our method is not limited to human body description and is also well suited to deal with any subjects. This particularity is due to the high-level description of the user-based constraints and to the fact that the sample-based pose definition is fully automatic.

II. Pose Description

In this section, we introduce a pose description based on local relations between 3D body part positions. We first define the input data required for our pose description, then we present the relations on which our description is based. Finally,

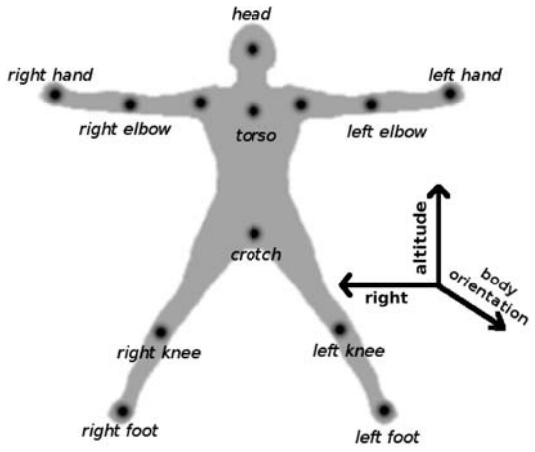


Fig. 1. Illustration of pose information.

we propose a technique to store these relations in order to obtain our pose description.

A. 3D Subject Information

In the case of human body, we consider the 3D positions of the following body parts: head, torso, crotch, hands, elbows, shoulders, knees and feet. Such information can be provided from multi-camera systems (e.g. using the methods proposed by Caillette and Howard [8], Michoud et al. [9] or Menier et al. [10]) or by specific devices like Kinect, with an appropriate library. We also consider that the floor normal orientation is known, such we can compute a set of orthogonal vectors, providing intuitive orientations for relative position definitions:

- the *altitude* vector, directly provided by the floor normal.
- the *right* vector, defined using the position of the two shoulders, and parallel to the floor.
- the *body orientation* vector, defined as being orthogonal to the two other vectors.

This input information is depicted in Fig. 1.

B. Intuitive Relations

A standard way to define feature description consists on a set of measures between parts, i.e. distances or angles, as presented by Kleinsmith and Bianchi-Berthouze [11].

Another approach, proposed by Müller et al. [12], consists in describing the pose using boolean features expressing geometric relations between certain body parts. For example, whether the right foot lies in front of or behind the plane spanned by three others specific body parts.

Our approach belongs to this latter, since we use boolean features for the description. However, two points differentiate our approach from the Müller et al. method. First, instead of selecting a subset of the geometric relations, our description include all of them, in order to be able to search for any defined relation between parts. The second difference concerns the kind of relations used for the description. In our case, we will use relations which can be easily expressed in natural language, in order to efficiently translate requests of users into conditions on relation set.

Such relations are more intuitive than measures to describe a pose and closer to a natural language: for example, it is more usual and easy to say that a subject has a hand *above* the head, rather than the subject has a hand at a distance n from the head and the angle a between the normal on the floor and the vector head-hand.

Naturally, one can argue that the use of relations, instead of measures, induces lower accuracy. However relations present some advantages, such as to be scale invariant, contrary to distances. Furthermore, according to the motion capture method used and to the input images resolution, a relation estimation between the body parts is less subject to noise than a measure. Considering these facts, the difference of accuracy between the use of measures and the use of relations is not so problematic.

In the following parts, we propose two kinds of relations: axis based relations, which describe the position of a part in regard to another and betweenness relations, which describe the position of a part, in regard to the position of two others.

III. Axis Based Relations

An axis based relation is a binary relation between two points in regard to their orthogonal projection on a given axis.

A. Intuitive Axis Definition

For the needs of our application, we will define three vectors forming an orthogonal base, and representing three intuitive orientations for human pose description. These vectors are defined using both the floor orientation and the body orientation. For each axis X defined in this paper, we denote the unitary vector of this axis by $\overrightarrow{\mathbf{X}}$.

The vertical axis is a very usual axis, measuring the distance to the considered point from the floor. This axis (denoted u) is a line normal to the floor, that has as origin an arbitrary point of the floor and which orientation is from the floor to the ceiling. The vector $\overrightarrow{\mathbf{u}}$ is depicted by the arrow "altitude" in Fig. 1.

In order to defined the right and left directions, we can consider the axis r supported by the projection on the floor plan of the vector defined from the left shoulder to the right shoulder. By definition, $\overrightarrow{\mathbf{u}}$ and $\overrightarrow{\mathbf{r}}$ are orthogonal. The vector $\overrightarrow{\mathbf{r}}$ is depicted by the arrow "right" in Fig. 1.

Finally, we can define the axis f representing the orientation of the torso, such that $\overrightarrow{\mathbf{f}} = \overrightarrow{\mathbf{u}} \times \overrightarrow{\mathbf{r}}$. The vector $\overrightarrow{\mathbf{f}}$ is depicted by the arrow "body orientation" in Fig. 1.

B. Positional Axis Based Relation

For two distinct projected points belonging to an axis, one of them will always have a greater measure than the other. Thus, for a given axis X, the *positional axis based relation* for two points p_1 and p_2 will be in the form: "the measure of the orthogonal projection of p_1 on X is greater than the

measure of the orthogonal projection of p_2". We denote this relation by $X(p_1, p_2)$.

Considering the three axis defined at the beginning of the section, we can introduce three positional axis based relations. For two points p_1 and p_2:

- p_1 *is above* p_2 if and only if $u(p_1, p_2)$.
- p_1 *is on the right of* p_2 if and only if $r(p_1, p_2)$.
- p_1 *is in front of* p_2 if and only if $f(p_1, p_2)$.

C. Distance on Axis Based Relation

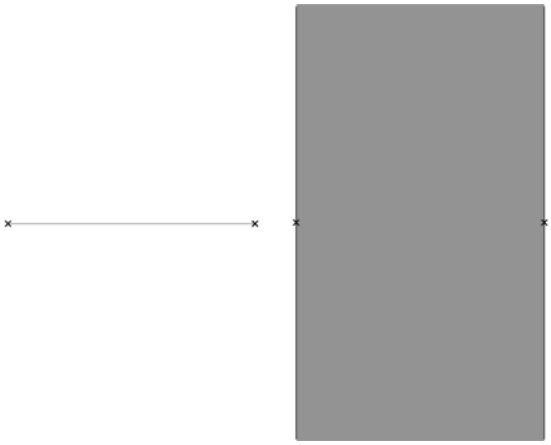
In order to improve the accuracy, we propose to add relations, in order to differentiate when two parts are close or far in regard to an axis. Let us consider a distance δ, which is considered as constant for each human body and equal to half the distance between the two shoulders. For two points p_1 and p_2, and for an axis X, p_1 and p_2 are *far from each other on* X if the distance between their projection on X is higher than δ. We denote this relation by $X_\delta(p_1, p_2)$.

IV. Betweenness Relations

Betweenness relations are ternary relations between three points, considering that one of them is *between* the two others. For robustness purposes, the main problem is to find a definition of betweenness relation which is not too restrictive nor too permissive. Indeed, an efficient betweenness relation would provide at least one positive response for a large panel of configurations of three points and at most one positive response for a set of three points.

A usual definition of a betweenness ternary relation is based on collinearity [13]: a point p_3 is said to be between two other points p_1 and p_2 (denoted by $B_c(p_1, p_2, p_3)$) if p_3 belongs to the segment p_1p_2. However, this definition is too restrictive. This relation is illustrated in Figure 2(a). On the other hand, we could say that a point p_3 is between p_1 and p_2 (denoted by $B_p(p_1, p_2, p_3)$) if its orthogonal projection on the line (p_1p_2) is between p_1 and p_2. In this case, the problem is that there exist triplets (p_1, p_2, p_3) such that each point is between the two others (for example, the three vertices of an equilateral triangle). Thus, $B_p(p_1, p_2, p_3)$ is too permissive. This relation is illustrated in Figure 2(b).

We propose a definition which is not too restrictive and which gives at most one possibility of "betweenness" for three points p_1, p_2, p_3: consider the unique ball $\mathcal{B}$ defined by the unique ball with diameter p_1p_2 minus the unique sphere with diameter p_1p_2 (i.e. the open ball with diameter p_1p_2). The point p_3 is between p_1 and p_2 (denoted by $B_b(p_1, p_2, p_3)$) if and only if $p_3 \in \mathcal{B}$. This relation is illustrated in Figure 3. An other formulation is that p_3 is between two points p_1 and p_2 iff the angle between $\overrightarrow{p_3p_1}$ and $\overrightarrow{p_3p_2}$ is greater than $\frac{\pi}{2}$. This definition is easy to use in our context and provides a unique result for a wide set of configurations of three points. Indeed, all the configurations where the 3 points are the vertices of obtuse or right triangles lead to a unique result. In the case of an equilateral triangle, none of the three vertices is considered as between the two others.



(a) B_c (b) B_p

Fig. 2. All points in the grey area are between the two points represented by black crosses: (a) in the sense of B_c, and (b) in the sense of B_p.

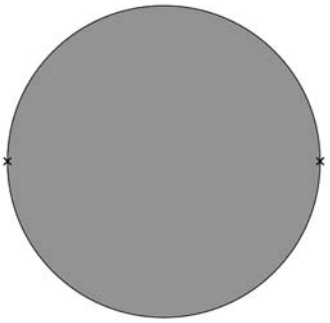


Fig. 3. All points in the grey area are between the two points represented by black crosses, in the sense of B_b.

In the following parts, $B_b(p_1, p_2, p_3)$ will be the only ternary relation we will use, hence we can shorten it as $B(p_1, p_2, p_3)$.

V. Comparison of Pose Descriptions

In this section, we propose several similarity measurements between two pose descriptors. These similarities between poses can be defined with the computation of a distance between the two sets of their relations.

A. Distance between Poses

In order to estimate the similarity between two poses, we propose to use the Jaccard distance between their relation sets. The Jaccard distance between $\mathcal{R}_1$ and $\mathcal{R}_2$ is computed as following:

$$d(\mathcal{R}_1, \mathcal{R}_2) = 1 - \frac{|\mathcal{R}_1 \cap \mathcal{R}_2|}{|\mathcal{R}_1 \cup \mathcal{R}_2|}$$

Notice that such distance is normalized: $d(\mathcal{R}_1, \mathcal{R}_2) = 0$ if $\mathcal{R}_1$ and $\mathcal{R}_2$ are equal, and $d(\mathcal{R}_1, \mathcal{R}_2) = 1$ if they are disjoint.

Some examples of poses[1] at different distances from a sample pose are depicted in Fig. 4.

[1]Data sets from MIT (http://people.csail.mit.edu/drdaniel/mesh_animation/)

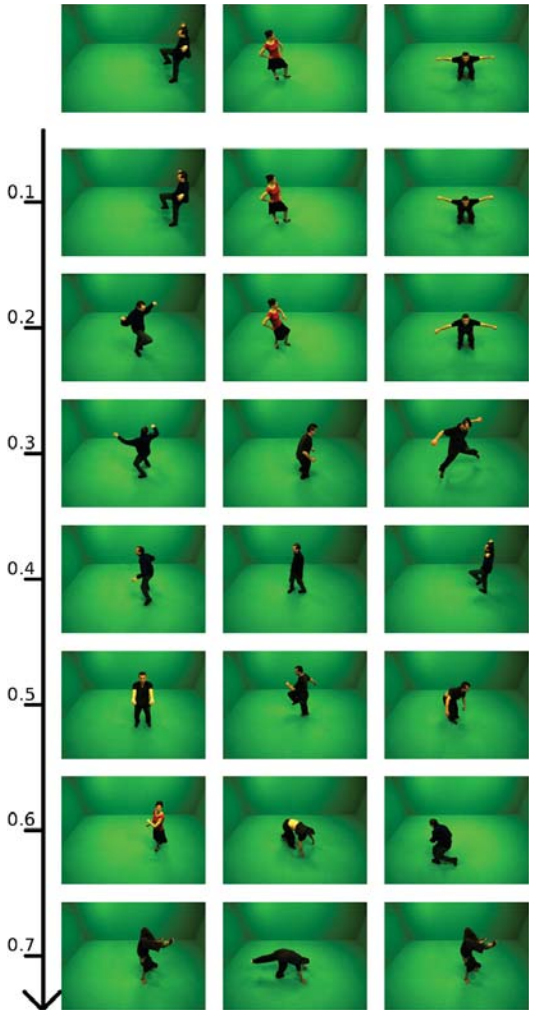


Fig. 4. Examples of poses in regard to the Jaccard distance to the sample (top image).

B. Distance Related to Specific Part

In some cases, it can be useful to evaluate the distance between a specific sub part of relation sets. For example, if we want to evaluate only the similarity of the upper part of two poses, independently from the lower part.

In that situation, we can use a specific pose descriptor containing all the relations we are looking for. In the case of the upper part example, it will consist in the set $\mathcal{R}_{up}$, such for any relation a, $a \in \mathcal{R}_{up}$ if and only if a is a relation involving a body part which is not in the upper part (e.g. a foot).

Using this set, we can define a distance between $\mathcal{R}_1$ and $\mathcal{R}_2$ in regard to $\mathcal{R}_{up}$ as following:

$$d_{up}(\mathcal{R}_1, \mathcal{R}_2) = d(\mathcal{R}_1 \cap \mathcal{R}_{up}, \mathcal{R}_2 \cap \mathcal{R}_{up})$$

C. Inclusion Measurement between Sets

For certain applications, it is useful to detect if a description is contained in another one. For example, we can imagine some applications using intersection of several poses descriptors, or their union. Thus, we want to be able to detect if the intersection is included in another pose (in the first case), or if a pose is included in the union (in the second case).

For this purpose, we can use a variation of Jaccard distance, the *inclusion measurement*, in order to measure the inclusion of one of the pose descriptor in the other. This measurement is defined as following:

$$im(\mathcal{R}_1, \mathcal{R}_2) = 1 - \frac{|\mathcal{R}_1 \cap \mathcal{R}_2|}{min(|\mathcal{R}_1|, |\mathcal{R}_2|)}$$

Contrary to the Jaccard distance, the inclusion measurement is not a distance, as the triangular inequality is not preserved.

VI. Syntax for Pose Description

A significant property of the proposed relations is that they are easily understandable by users and close to those used in a usual description in a natural language. Thus, it is easy to convert a natural description of relative positions of body parts into conditions on the relation set $\mathcal{R}$, e.g.:

- the subject has the left hand above the head: $u(hand_l, head) \in \mathcal{R}$
- the subject has the head between its hands: $B(hand_l, head, hand_r) \in \mathcal{R}$

A. Implicit Relations

In addition, we can obtain more complex relations than those contained in the descriptor, using combinations of them. For two body parts p_1 and p_2:

- p_1 *is far on the right of* p_2 if and only if $r(p_1, p_2) \in \mathcal{R}$ and $r_\delta(p_1, p_2) \in \mathcal{R}$. We denote this relation by $R(p_1, p_2)$.
- p_1 *is far above* p_2 if and only if $u(p_1, p_2) \in \mathcal{R}$ and $u_\delta(p_1, p_2) \in \mathcal{R}$. We denote this relation by $U(p_1, p_2)$.
- p_1 *is far in front of* p_2 if and only if $f(p_1, p_2) \in \mathcal{R}$ and $f_\delta(p_1, p_2) \in \mathcal{R}$. We denote this relation by $F(p_1, p_2)$.
- p_1 *is close to* p_2 if and only if $f_\delta(p_1, p_2) \notin \mathcal{R}$, $r_\delta(p_1, p_2) \notin \mathcal{R}$ and $u_\delta(p_1, p_2) \notin \mathcal{R}$. We denote this relation by $C(p_1, p_2)$.

B. Syntax Definition

We have now defined eleven relations: the seven contained in the descriptor and the four explicit ones. We can also defined their opposite very simply. Combining such relations, which can be directly convert from natural language, with boolean operators, like AND, OR, XOR, and NOT, we obtain a syntax providing a good description of any kind of pose and still easily understandable by the user. In order to lighten the writing of the syntax, we propose to do the following replacements for any relation x:

- $x(p_1, p_2)$ instead of $x(p_1, p_2) \in \mathcal{R}$
- $NOT\ x(p_1, p_2)$ instead of $x(p_1, p_2) \notin \mathcal{R}$

Here are some examples of pose definitions with this syntax:

- the head is between the hands, at closely the same altitude:

$$B(hand_l, head, hand_r)\ AND$$
$$u_\delta(hand_l, head)\ AND\ u_\delta(hand_r, head)$$

- only one hand above the head:
  $u(hand_l, head)\ XOR\ u(hand_r, head)$
- at least one hand close to the torso:
  $C(hand_l, torso)\ OR\ C(hand_r, torso)$

VII. Application for Pose Recognition and Results

An interesting application for our pose description is the pose recognition for human-machine interfaces or surveillance applications, as the pose descriptor as well as the comparisons can be easily performed in real time. Indeed, for an implementation in C++ running on an average computer (dual core 2.8GHz), the computation of a pose descriptor is done in less than two milliseconds. Depending on the structure used for the representation of the relation sets, between 10000 and 100000 distance computations between two poses can be performed in one second, and over 10000000 conditions on a pose descriptor can be tested.

We propose two ways to define the poses which have to be detected:

- using the syntax defined in Sec.VI
- using a set of samples

In our experiments, all the samples and detections are done using Kinect.

A. Pose Recognition from Syntax

In order to test the efficiency of the syntax description, we defined a set of simple descriptions, using those introduced in Sec. VI-B and the following ones:

- two hands far in front:
  $F(hand_r, torso)\ AND\ F(hand_l, torso)$
- right foot up:
  $u(foot_r, knee_l)$
- crossed arms:
  $R(hand_l, hand_r)$

some results of pose detection using these descriptions are shown in Fig. 5. The accuracy of the results depend on different parameters: the accuracy of the motion capture method, its robustness to quick movements, and the precision of the floor normal definition. In our case, the use of Kinect provides good results in regard to the quality of the motion capture. Of course, sometime the tracking failed resulting in false positives (around 10 percent of the utilization time), but in other cases the detection is working efficiently. The other sensitive point is the positioning of the device, providing an inaccurate floor normal, resulting on some lack of accuracy on the altitude axis based relations.

Fig. 5. Examples of poses detected using syntax description.

B. Pose Recognition from Samples

An alternative to the description by syntax is the description by samples: we provide in input a set of poses having in common the pose specificities we want to detect. The intersection of all their relation sets preserves only their common part, and thus what we want to detect. We can combine the use of samples with a restrictive relation set, for example in order to take into consideration only the upper part of the body, or only a sub set of relations. Figure 6 shows the set of poses used for our tests of pose recognition by samples. The detection is performed with the inclusion measurement between the current pose and the intersection of the samples: if it is equal to 0, it involves that the current pose contains the same common part that all the samples, and thus what we want to detect.

Figure 7 show the detections of recorded poses (those illustrated in Fig. 6) obtained for different poses.

The recognition from sample is more robust than the one from syntax, due to the fact that the samples are captured with the same device/method which is then used for the detection. However, it is more restrictive, as it cannot be used to described relations like “at least one of the arms” or “only one of the feet”.

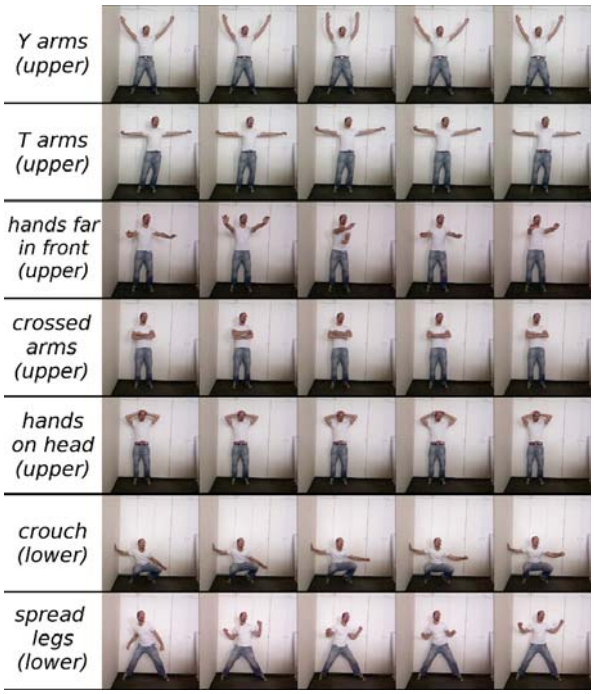


Fig. 6. Sample poses used for detection. Each row represents a different pose to detect, with its name at the left and the restriction set used into bracket: upper or lower part of the body

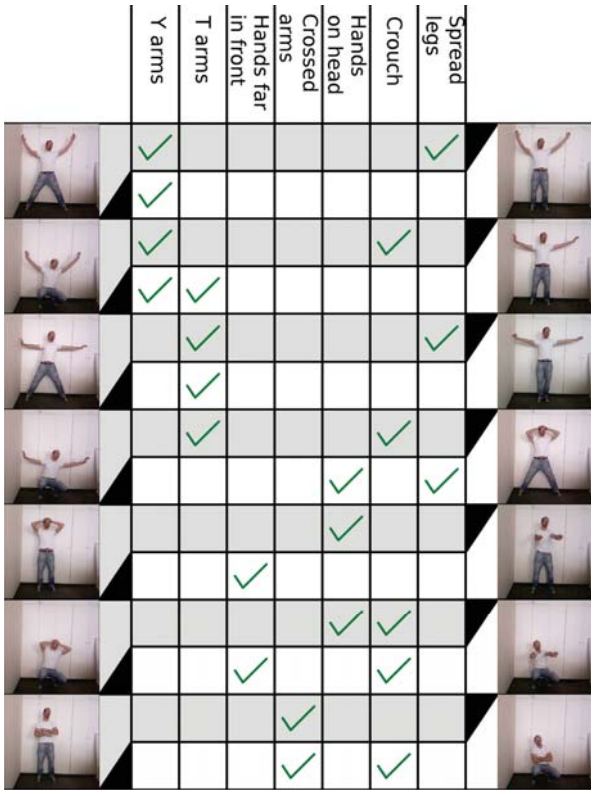


Fig. 7. Examples of poses detected using sample description.

VIII. CONCLUSION

This paper presents a new approach to describe the 3D pose of a subject using local and simple relationships between the subject parts, namely the axis-based and betweenness relations. We show that using a set of these relations is an efficient tool to represent a pose as well as to compute a similarity between two poses. Indeed, we present a bunch of rules that can describe a 3D pose. These rules can be defined intuitively since they look like a natural language. We also show that our method is well suited to compare different 3D poses. We introduce a similarity measure to express this difference. Our results shows that the pose computation and identification run in real-time.

As a future work, we plane to extend our method to handle more complex gesture recognition. A possible approach would be to take a leaf out of the papers of Lv et al. [14] or Weinland et al. [15] that define a path of state to recognize human motion patterns. As in this paper, we want to keep real-time processing to ensure the compatibility of our method with some human machine interfaces.

REFERENCES

[1] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[2] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 3, no. 37, pp. 311–324, 2007.

[3] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 1, pp. 13–24, 2009.

[4] A. Agarwal and B. Triggs, "3d human pose from silhouettes by relevance vector regression," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 882–888, 2004.

[5] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1593–1600, 2009.

[6] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," *Computer Vision, IEEE International Conference on*, vol. 2, p. 750, 2003.

[7] J. Sullivan and S. Carlsson, "Recognizing and tracking human action," *In European Conference on Computer Vision (ECCV 2002)*, pp. 629–644, 2002.

[8] F. Caillette and T. Howard, "Real-time markerless human body tracking with multi-view 3d voxel reconstruction," *in the British Machine Vision Conference*, vol. 2, pp. 597–606, 2004.

[9] B. Michoud, E. Guillou, H. Briceno, and S. Bouakaz, "Real-time marker-free motion capture from multiple cameras," *in the IEEE Eleventh International Conference on Computer Vision*, pp. 1–7, 2007.

[10] C. Menier, E. Boyer, and B. Raffin, "3d skeleton-based body pose recovery," *in the Third International Symposium on 3D Data Processing, Visualization and Transmission*, pp. 389–396, 2006.

[11] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," *in International Conference on Affective Computing and Intelligent Interaction*, pp. 48–58, 2007.

[12] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," in *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 677–685.

[13] M. Greenberg, "Euclidean and non-euclidean geometries: Development and history," *WH Freeman*, 1993.

[14] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[15] D. Weinland, F. Grenoble, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," *in Proc. IEEE Conf. Computer Vision*, pp. 1–7, 2007.