Grasp Recognition and Manipulation with the Tango

Paul G. Kry[1,2,3] and Dinesh K. Pai[1,2]

[1] Department of Computer Science, Rutgers University,
New Brunswick, NJ, USA.
[2] Department of Computer Science, University of British Columbia,
Vancouver, BC, Canada.
[3] EVASION, INRIA,
Montbonnot, France.

Summary. We describe a novel user interface for natural, whole hand interaction with 3D environments. Our interface uses a graspable device called the Tango, which looks like a ball but measures contact pressures on its surface at 256 tactual elements (taxels) at a high rate (100 Hz). The acceleration of the device is also measured. The key idea is to use this information to recognize the shape and movement of the user's hand grasping the object. This allows the user to interact with 3D virtual objects using a hand avatar. The interface provides passive force feedback, and is easier to use than interfaces that require wearing gloves or other sensors on the hand. We describe a rotationally invariant matching algorithm for recognizing the hand shape from examples of previous interaction collected with motion capture. We also describe examples of 3D interaction using our system.

1 Introduction

Programming robots by demonstration has long been a dream of robotics, but it has remained elusive for complex tasks involving grasping and manipulation. This is, in part, due to the difficulty of simultaneously capturing the configuration of the user's hand *and* the intended contact forces. In addition, one would like the manipulandum to be simple and easy to use, without requiring cumbersome motion capture equipment or instrumented gloves that distract from the task at hand. One possible option is to use a manipulandum with a pressure sensitive skin and inertial sensors for quickly recognizing the shape and movement of the user's hand grasping the object, with visual feedback provided by 3D virtual environment. Such a system would make it much easier to program interaction with 3D objects.

In this paper, we describe one way to achieve this type of natural interface using a device called the Tango. The Tango, whose name is derived from the word "tangoreception" (meaning pertaining to the sensation of touch), is a ball that fits conveniently in the hand. There are 256 pressure sensors on the device's surface and a 3-axis accelerometer within. We describe a new rotationally invariant algorithm for recognizing hand configuration from pressure on the surface of the Tango, by using examples of previous interaction collected with motion capture. The recognition method is sufficiently fast for interactive manipulation. We also describe examples of 3D interaction using our system, in which the user interacts with 3D virtual objects using a hand avatar.

2 Related Work

Our previous work on the Tango [11] described the design of the Tango device, and presented a simple method for grasp tracking. In this paper, we focus on recognizing realistic hand shapes and using this and other input from the Tango for 3D interaction.

Glove-based interfaces are currently the most common whole-hand user interfaces [4, 14], though computer vision has also been used (e.g.,[10]). The lack of force feedback is an important limitation with these interfaces. Several devices address this problem by providing active force feedback [2, 5]. However, whole hand force feedback is expensive and complex; passive force feedback via a tangible object such as a ball is often sufficient [16, 6].

Reconstruction of full body posture from foot pressure data [15] is a similar problem but requires a different solution because the latency requirements are more severe for manual interaction than for animation. Previous work on grasp recognition includes [1], which uses both forces and the hand shape to classify grasps for robotic programming by demonstration.

3 Technical Approach

We recognize the user's hand configuration by rotationally invariant comparisons of pressures on the Tango with previous training measurements that capture both the pressures and the actual 3D hand shapes during manipulation. This section explains our method in three parts: clustering and identifying fingers, grasp hashing, and grasp identification.

3.1 Clustering and Identifying Fingers

We first cluster taxels (tactual elements) for different contacts to determine the number of fingers that are involved in a grasp, and compute a pressure centroid and a total pressure for each cluster.

At each activated taxel, we search its four directly connected neighbours (east, west, north, and south) and perform a merge if any of the neighbours are activated. In addition, we also check two additional taxels along the same meridian (the second taxel to the north and the second taxel to the south). Since variation in the sensitivity of taxels can result in taxels that do not activate during light grasps, this allows for clusters with a vertical gap (for example, see the bottom left corner of Figure 1).

In the case of three-finger grasps, we have also explored the use of heuristics to identify which finger is responsible for each cluster. The thumb cluster is almost always identifiable as the cluster with the greatest total pressure. Assuming the Tango is grasped from above with the right hand, then starting at the thumb cluster and travelling westward along the surface of the Tango, we identify the next cluster with the index finger and the next following with the middle finger. Furthermore, we restrict the search for the middle finger to meridians that are within 45 degrees of the meridian opposite the thumb cluster. This avoids identifying spurious single taxel clusters (caused by noise) as finger plants. The alternative is the arbitrary removal of single taxel clusters from consideration. Results of these finger heuristics can also be seen in the left hand side of Figure 1, where thumb, index, and middle finger clusters are coloured red, green, and blue, respectively. Observe that the thumb heuristic is not sufficient to disambiguate the two finger grasp, but this case can be handled by taking into account continuity with previous grasps.

3.2 Grasp Hashing with Spherical Harmonics

Inverse kinematics and the location of finger plants (as identified by heuristics) could be used to produce grasp configurations; however, the inverse kinematics problem is underconstrained. Instead, we use example data to resolve the redundancy. Using previously collected example data, we associate a distribution of natural hand configurations with observed pressure measurements. A plausible hand shape can then be selected from the distribution. In this manner, we can infer the pose of *all* fingers from the pressure generated by just those fingers that are in contact.

We perform rotationally invariant comparisons, so that identical pressure distributions applied at different orientations will match. Similar to the work of Kazhdan, et al. on shape matching [7], our spherical pressure functions can be transformed into rotationally invariant features.

We first project the pressures p_{ij} on to real-valued bases y_l^m derived from spherical harmonics and sampled at the taxel locations. The coefficients are

$$a_l^m = \sum_{i,j} y_l^m(\theta_j, \phi_i)\ p_{ij}, \tag{1}$$

where θ_j and ϕ_i provide the polar and azimuth angles of the taxel centers, and p_{ij} is the pressure of the taxel located on meridian i and parallel j. We precompute y_l^m since the taxel locations are fixed. The pressure function in the spherical harmonic basis is a frequency-limited smoothly varying representation.

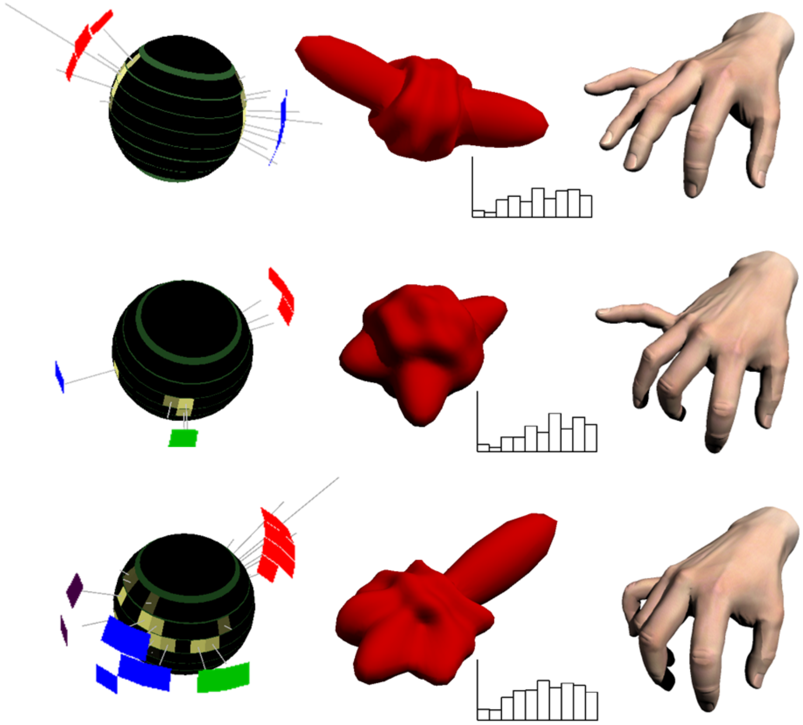

Fig. 1. Tango data, clusters, spherical harmonics, and hand pose

We use 10 frequencies, $f = 10$, in our spherical harmonic basis, which corresponds to a total of 100 basis functions, since there are $2l - 1$ functions at each integer frequency l. Note that 10 is a user-selected parameter; we need $f = 16$ to make Equation 1 invertible. With our smaller value of f, Equation 1 is a projection and acts like a low pass spatial filter. Given the size of fingerpads in comparison to taxel areas, we believe the omission of the higher frequencies is reasonable.

The sum of the energies (ℓ^2 norm) at each of the first f frequencies produces a histogram $x = (x_0, \cdots, x_f)^T$,

$$x_l = ||a_l||, \tag{2}$$

where $a_l = (a_l^{-l}, \cdots, a_l^l)^T$ is the vector of coefficients at frequency l. This histogram can be thought of as a feature vector, fingerprint, or hashing of the pressure function, with built-in rotational invariance. A key feature of this hash function is that it is locality-preserving. Specifically, a set of similar grasps result in similar histograms, while a set of similar histograms correspond with subsets of similar grasps.

Because our example data consists only of pressures produced by a hand and does not contain arbitrary pressure images, there exists a fair amount of

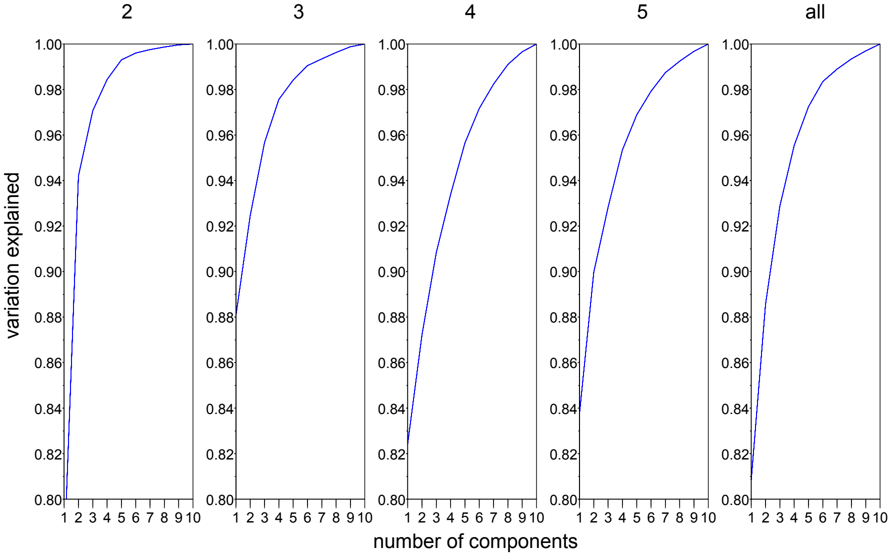

Fig. 2. Variation explained when using different number of components, shown for 2, 3, 4, and 5 finger precision grasps, and for all data.

redundancy in the histograms. Principal component analysis (PCA) of the example data energy histograms provides a smaller orthogonal basis in which we can compare measurements. Projecting the histograms into a truncated PCA space reduces the sparsity of previously collected data (and lowers memory requirements). It also lets us compute more meaningful distance comparisons by discarding dimensions that contain only noise while boosting the contribution of important dimensions with small variance. Previous work has shown that final grasp postures are well approximated by only a few principal components [13]. Likewise, our measured variations in hand shape (similarly the pressure distribution and corresponding histogram) are well approximated by a lower dimensional subspace, especially considering that the user's hand is constrained to be grasping an object of fixed shape, the Tango. Figure 2 shows that only a few components are necessary to explain 90% of the variation. We project each histogram into a previously computed truncated PCA space to produce a d-dimensional vector representing the current grasp shape (we used $d = 6$ in our experiments). We refer to these vectors as *pressure hashes* and use comparisons of them for grasp identification as described below.

3.3 Grasp Identification

We acquire example data that includes *both* grasp pressures *and* hand configuration, measured using a Vicon motion capture system. For run-time grasp identification, we use the pressure hash to find the k-nearest (Euclidian dis-

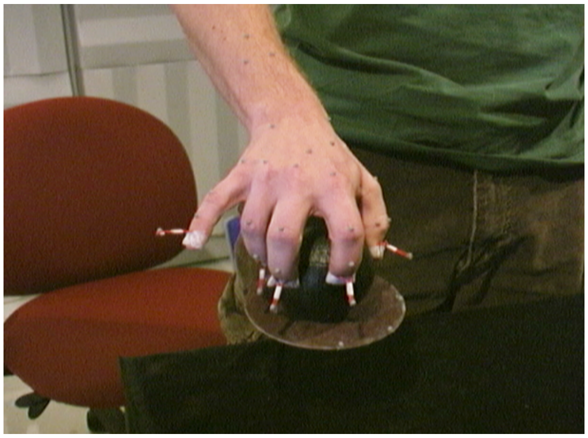

Fig. 3. Example four finger grasp collected with motion capture. Fingertips have markers on "stilts" to reduce occlusion during grasps.

tance) neighbours in the previously computed data. For this we use a bounding hyper-sphere tree constructed with the method described by [12] but extended to arbitrary dimension. Recall that building a tree of data in PCA coordinates lets us easily compute Mahalanobis-like distances in different truncated spaces by simply summing fewer terms in our ℓ^2 distance computation. The bounding sphere tree is still valid for truncated spaces, though possibly less efficient.

Each of the k neighbours for the current pressure measurement has a corresponding hand configuration, which we compare with our current hand configuration using a weighted Euclidean distance. The weighted distance metric allows us to ignore the position and orientation of both the forearm and wrist. Overall, our method works much like a simplified particle filter tracker [3]. The closest hand configuration among the k-nearest pressure-hash neighbours becomes the proposal configuration. Note that if there is no pressure observed on the Tango, then we can infer nothing about the hand shape. In this case, we use a previously selected rest pose configuration for the proposal.

4 Results

We used the Tango [11] in our experiments (see Figure 4). It produces an 8x32 tactual image with 8 bits per taxel, and a 3-axis acceleration reading, at 100 Hz. Filtering techniques for the raw data are described in [11]. Figure 1 (left column) shows examples of the initial pressure clustering, where thumb, index, and middle finger clusters are coloured red, green, and blue, respectively.

To build our example data set, we acquired synchronized motion capture of the Tango position and orientation, hand configuration, taxel pressures,

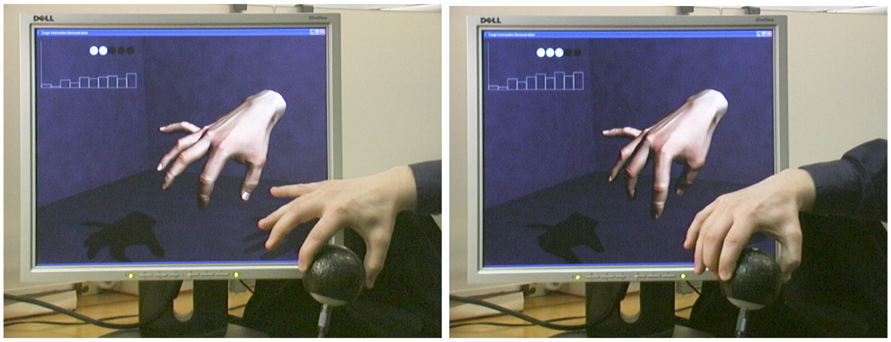

Fig. 4. Grasp approximation results

and Tango accelerations. We used a 6-camera Vicon motion capture system (Vicon Peak, Lake Forest, CA) to track small retro-reflective markers on a subject's hand (see Figure 3). Interactions with different numbers of fingers were considered separate "conditions". In total, approximately 10 minutes of capture data was acquired at 60 Hz. For each condition, we compute a separate PCA space of the energy histograms of surface pressure samples.

Figure 1 shows three example data points from the two, three, and four finger trials. The left column shows raw taxel data (pressure magnitudes shown by lines emanating from activated taxels, shaded yellow) with clusters shown in unique colours displaced from the surface. The center column shows the spherical harmonic representations for the pressure data and its 10-frequency energy histogram. The right column shows the corresponding synchronously captured hand pose.

The recognition algorithm using nearest neighbour searches are very fast because of the simple bounding volume test, combined with small tree depths. Our deepest tree has 19 levels for about 9500 data points.

To improve performance, we use the finger count from clustering to restrict our search for proposals to only the example data containing grasps with the same number of finger plants. Figure 4 shows our approximation result for a two-finger and three-finger grasp.

5 Experiments

We have developed a small virtual world in which we can explore the performance of positioning, orienting, and object interaction tasks. Figure 5 left shows a snapshot of the user's view of the world. Note that grasping in our demonstration is iconic, though we could use simulation to bring the hand into contact with the object [8]. Positioning and targeting the hand using only accelerometers is difficult, and could be improved by addition of gyroscopes that are now readily available. Nevertheless, we implemented a simple positioning

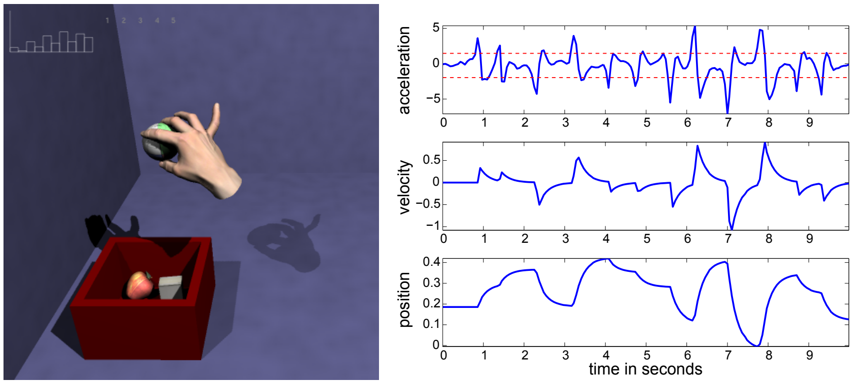

Fig. 5. Left, a screen shot from the Tango interaction demonstration. Right, a graph of Tango vertical position control from acceleration.

interface that uses the measured attitude for velocity and position control for experimentation (see Figure 5 right).

We can also use the grasp information for mode selection. Specifically, the number of fingers used in a grasp, as identified by clustering, provides a reliable method of mode selection. Virtual or free form buttons can also be implemented this way. In our initial experiments, we tried assigning different fingers to different virtual buttons, but it is difficult to control the pressure of one finger independently of the others because the user's fingers must satisfy a force closure property on the Tango to maintain a stable grasp. Instead, the number of fingers used in a grasp can determine the button number.

Using this interface, the user can grasp an object, rotate it in 3D, transport it to a different location, and place it, using hand movements analogous to those that would be used in a real setting. In the future, such tangible interfaces could be used for programming robots by demonstration or for model-based telerobotics [9].

6 Conclusions

This paper presents a novel user interface for 3D whole hand interaction using a new interface called the Tango. Hand shapes during grasping can be recognized from pressure distributions using rotationally-invariant feature matching and a collection of interaction examples collected with synchronized Tango and motion capture data. With this interface, the user receives passive haptic feedback while performing 3D interaction. In our experiments, the user is shown a hand avatar that mimics the shape and motion of the user's hand without the use of a glove.

6.1 Limitations and Future Work

We assume all grasps on the Tango are precision grasps, which simplifies the identification of number of fingers as the number of clusters; however, rotationally invariant pressure hashes show promise for correctly identifying the number of fingers and hand shape when all of our example data trials are combined into one. Furthermore, we expect our method extends to other grasp types, such as conforming and palmar grasps.

Acknowledgements

This work was supported in part by NSF grants IIS-0308157, ACI-0205671, and EIA-0215887, and the IRIS Network of Centres of Excellence. Thanks also to Maya software donated by Autodesk.

References

1. K. Bernardin, K. Ogawara, K. Ikeuchi, and R. Dillmann. A sensor fusion approach for recognizing continuous human grasping sequences using hidden Markov models. *IEEE Transactions on Robotics*, 21(1):47–57, February 2005.
2. Mourad Bouzit, Grigore Burdea, George Popescu, and Rares Boian. The Rutgers Master II–New design force-feedback glove. *IEEE/ASME Transactions on Mechatronics*, 7(2), June 2002.
3. Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.
4. Immersion Corporation. CyberGlove.
5. Immersion Corporation. CyberGrasp.
6. B. Insko, M. Meehan, M. Whitton, and F. P. Brooks Jr. Passive haptics significantly enhances virtual environments. Technical report, Computer Science Technical Report 01-010, University of North Carolina, Chapel Hill, NC, 2001.
7. Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In *Proceedings of the Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 156–164. Eurographics Association, 2003.
8. Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *ACM Transactions on Graphics*, 25(3):872–880, 2006.
9. John E. Lloyd, Jeffrey S. Beis, Dinesh K. Pai, and David G. Lowe. Programming contact tasks using a reality-based virtual environment integrated with vision. *IEEE Transactions on Robotics and Automation*, 15(3):423–434, June 1999.
10. Shan Lu, Gang Huang, Dimitris Samaras, and Dimitris Metaxas. Model-based integration of visual cues for hand tracking. In *WMVC*, 2002.
11. D. K. Pai, E. W. VanDerLoo, S. Sadhukan, and P. G. Kry. The Tango: A tangible tangoreceptive whole-hand human interface. In *in Proceedings of WorldHaptics (Joint Eurohaptics Conference and IEEE Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems)*, Pisa, Italy, March 18-20, 2005.

12. S. Quinlan. Efficient distance computation between non-convex objects. In *IEEE International Conference on Robotics and Automation*, pages 3324–3330, 1994.
13. M. Santello, M. Flanders, and J. Soechting. Postural hand synergies for tool use. In *The Journal of Neuroscience*, 1998.
14. Sarcos. http://www.sarcos.com/telerobotics.html.
15. KangKang Yin and Dinesh K. Pai. FootSee: an interactive animation system. In *Proceedings of the ACM SIGGRAPH Symposium on Computer Animation*, pages 329–338. ACM, July 2003.
16. S. Zhai, P. Milgram, and W. Buxton. The influence of muscle groups on performance of multiple degree-of-freedom input. In *Proceedings of CHI '96*, pages 308–315, 1996.