Advances in the Study of Hand Gesture Recognition Systems for Human Computer Interaction

P. Rodrigo Díaz-Monterrosas, Rubén Posada-Gómez, and Albino Martínez-Sibaja

Tecnológico de Orizaba, División de Estudios de Posgrado e Investigación, Orizaba, Veracruz, México
diaz.monterrosas@gmail.com, pgruben2@hotmail.com, albino_mx@yahoo.com

Abstract. Getting three-dimensional pose and orientation of parts of the body observed by one or more cameras is of great theoretical interest and widely applicable. Usually, computing devices interaction is accomplished by means of a mouse and a keyboard or by touching the screen, but otherwise, human beings relate to their surrounding world using hands, body, and voice in most of their daily activities, therefore, development of more natural and intuitive techniques for interacting with a variety of user interfaces is critical. In this paper, a review of recent research efforts in Human Computer Interaction (HCI), specifically in hand gesture recognition, is performed, analyzing the state-of-the-art methodology and discussing some important issues about.

Keywords: Hand Gesture Recognition, Human Computer Interaction, Image Processing, Computer Vision.

1 Introduction

Analyzing the different techniques used in literature to achieve location, tracking, description, and object recognition, has led to the development of tools that tend to improve robustness and naturalness in handling HCI devices to be fully functional in real world. Taking hands, face, body, voice, or even the eyesight as objects of study, research evolves allowing Douglas Engelbart's augmentation dream to increasingly become tangible, and disciplines such as Artificial Intelligence (AI) whose philosophy since its beginning has been considered as opposed to HCI's vision, devote part of its efforts to study and try to simulate human communication processes with the same purpose: interact in a friendlier way with machines. Among main achievements of these disciplines, Automatic -visual, speech, gesture, etc.- Recognition Systems designed to recognize and translate what they hear and/or see (voice, lips or body movements, facial gestures, hand signs or other objects movements) via microphones, video cameras or other sensors, could be mentioned.

Recently, new technologies such as RGB-D cameras, which have sensors to capture RGB images along with depth information of each pixel, have been taken

into account. These high precision cameras are capable of delivering high-quality three-dimensional information (color and depth) to understand the whole shape of an object [1]. For example, Microsoft® Kinect® technology allows players to enjoy video games simply by moving their bodies in front of a screen, the data taken as input from the device are the tracked body skeleton. The extension of this technology to individual finger 3D tracking is an active research area, having more complete information on user's hand pose would lead to grasping, pointing, and manipulation capable richer applications [2]. However, despite the growing amount of research in the area, there are still existing problems having a variety of theoretical and practical challenges.

The following section describes some actual marketed devices features, regarding their interaction drawbacks. The major current challenges in image-based gesture recognition are cited and later the resolution methods generally reported in literature are defined along with a description of some of the most representative recent projects. Finally, a future contribution and discussion to the described state-of-the-art is submitted.

2 Existing Devices, their Drawbacks and Health Impact

Nowadays, there are several devices with the adjective "smart" on them: digital still and/or video cameras with facial/smile recognition and tracking, automatic washing machines capable of weighing clothes and dosing optimum amount of water and detergent together with the estimation of washing time, computers and smartphones which light-up when passing the hand over them or with a voice command, or the new smart TVs, very trendy by additionally providing a web browser, access to several services, and likewise have the ability to perform all typical orders of their modern remote control by gestures or voices.

Even with such technological advances, these devices are far from being even slightly smart at least for now, because although they averagely meet what they promise, there are more than a few shortcomings in their performance, causing most of the time user prefers to disable smart features, and use traditional interaction ways (it might also be the case of the Leap Motion Controller, described as "rather limited" in [3]). For some smart TVs, for example, keeping alive gesture detection feature implies the stress of its continuous unwanted activation simply by raising arms or doing a hand sign to someone else in the room. Voice detection behaves similarly, activating itself even by the loudness of a movie dialog. As if this were not enough, when someone does want to use these features, orders hardly execute at first attempt, having the user to desperately repeat the gesture or key word once and again.

Clearly, at the moment of this study, there are still many challenges related to accuracy, speed, naturalness, and even usefulness of these devices, since although occasional user might be pleased or amazed, for regular user the continuous use of features such as gesture recognition, for example, could harm his health and comfort due to a phenomenon known as "gorilla arm syndrome", a problem that arises from continued use of the arms in the air, that is, without a place to

stand, causing a feeling of heaviness, fatigue, or discomfort (imagine a designer working eight hours on a gesture based software). On the other hand, it is true that this is a less significant problem if a similar disease suffered for decades due to extensive use of the mouse, known as "carpal tunnel syndrome", comes to mind (a hand and arm condition caused by the inflammation of a tendon in the wrist and triggering chronic pain), yet, it still remains one of the most used interaction device.

3 Current Challenges in Hand Gesture Recognition Area

Leaving aside negative issues on the development of image-based gesture recognition devices, there have been challenging factors since the beginning, such as the high dimensionality and speed of hand motion, while image capturing is performed through low resolution cameras, besides the diversity of existing cameras that hinder calibration or standard lighting conditions, the ambiguity of image elements identification due to its color uniformity, the similarity of fingers, the large number of degrees of freedom (DOF), or the absence of observations when parts of hands (or other object) obstruct each other. Searching for solutions, special hardware for motion capture has been used, such as magnetic tracking devices, bracelets with optical [4] or electromyographic [5] sensors, and visual markers placed on gloves [6] or the bare hands. Unfortunately, such methods require complex and expensive hardware, interfere with the observed scene, or add restrictions to user's pose, preventing their use in real world [7], not to mention their use for people with disabilities. Moreover, emerging projects addressing hand tracking interacting with objects, have increased the challenge, since although they can help to reduce the number of potential poses, there are limitations still being solved, such as the desirability of the hardiness and non-similar to skin color of the object.

4 Classification of Hand Gesture Recognition Methodology

From a historical perspective, starting with the development of articulated hands to explore issues related to grip and object manipulation at early 80s, a growing attention from a variety of disciplines to modeling, 3D simulation, tracking, and interpretation of hands and other body parts motion has been found (it is accepted that the study of hand tracking began with the Zimmerman et al "VPL Dataglove" [8]). Several applications have emerged from these studies, which have been useful to many science and technology areas, to name a few: medicine and biotechnology, robotics, computer animation, movies, e-commerce, and virtual reality.

At first systems, selection of key points (articular joints) was performed manually on the computer screen. Obviously those had serious restrictions, the most significant: selection subjectivity, process slowness, and calibration

system sensitivity. Such systems, typical of the 90s, are deprecated, and from the first decade of 21st century, finding semi-automatic methods with reflective, magnetic or infrared light sensitive optical markers is usual, allowing, when scanned, determine joint connections. Currently, research is mainly oriented to the development of processes that avoid the use of invasive elements.

Depending on the type of input, gesture translation approaches may vary. However, most techniques are based on key clues represented in a 3D coordinate system, by tracking their relative motion a gesture can be detected with high precision. Several methods have been used in the literature to estimate the hands pose, and have been classified by some researchers according to certain common properties; regarding the output completeness, Erol et al [9] describe them as "Partial hand pose estimation methods" which can be viewed as extensions of "Appearance-based methods" to provide information on continuous motion in navigation, handling or pointing; and "Full DOF hand pose estimation methods", which get all the kinematic parameters of the skeleton of the hand, such as joint angles and hand position or orientation for a full reconstruction of hand motion. The latter class is divided into: "Model-based tracking", which can be subdivided into methods using a single hypothesis and those who manage multiple hypotheses, and "Single frame pose estimation", that is, they are not committed to time coherence. Both full DOF hand pose estimation classes are addressed in [10].

4.1 Appearance-based Methods

This approach, also known as "Discriminative", uses classification or regression techniques directly into the image data. An offline training process is used to establish a nonlinear mapping (due to the different hand views) from the image feature space to a finite set of hand poses, depending on specific parts of the hand, such as palm or fingertips and their orientation. These methods process each image independently, but may be used with image sequences; they work well when recognition of a small well-known and distinct hand configuration set is required and are not recommended when there is free hand motion and high recognition accuracy is required. Velocity, offline training, computationally efficient online execution, low computational cost and hardware complexity, the requirement of a single camera, and generalization if training is suitable are some of their advantages. Their inherent disadvantages lie in the need for very large training data sets and that their accuracy and reduced number of hand recognition poses rely on those data, therefore requiring high degree of user intervention.

Recent research involving two-hand recognition introduce a new challenge if this approach is used, due to the fact that the offline training must include the combinatorial space of both hands configuration and the changes that different points of view cause in their appearance.

4.2 Model-based Methods

These methods are called "Generative" because they generate hypothetical 2D or 3D hand models and compare model projection to the observed images. This is done through an optimization problem whose objective function measures the discrepancy between the model key indicators and the observations, however, the optimization method should be able to evaluate the objective function at arbitrary points in the multidimensional space of model parameters, so the search must be carried online, causing a high computational cost, which is their major drawback, besides relying entirely on visual information available, usually provided by a multi-camera system. On the other hand, these aspects also imply their major strengths: there is no training need and they can easily be extended to any gesture recognition problem. If researcher decides to use this approach, dimensional reduction of the configuration space, efficient construction of realistic three-dimensional hand models, and development of quick and reliable estimating techniques would be interesting contributions [11].

The usual visual features to match are silhouettes, edges, shades, color, optical flow, and recently depth. Among the optimization techniques that have been proposed are, to name a few, belief propagation, particle swarm optimization, and local optimization, one of the first and still used because of its efficiency. Similarly, stochastic optimization techniques such as Kalman filter and particle filter have been used, the latter together with local optimization in [12]. In [13] and [14] linear subspaces are used to reduce the hand pose space.

In short, appearance-based methods allow fast processing with a loss of generality, whereas model-based ones give generality at a high computational cost.

Another classification is based on how partial evidence of individual rigid parts of an articulated object contributes to the final solution [15]:"Disjoint evidence methods" consider individual parts in isolation before evaluating them against observations, usually requiring less computational power but needing to handle explicitly part interactions (such as collisions and occlusions); in contrast, "Joint evidence methods", consider all parts in the context of full object hypotheses, their computational requirements are high, but part interactions do not represent much of a problem.

5 Current Research in Literature

Oikonomidis et al [11] introduce a model based multiple-view method to recover 3D position of the hand given by 27 geometric primitives that redundantly encode a 26 DOF 3D hand model. Observations are acquired from a static, pre-calibrated camera network, computing reference features for each acquired view based on skin color and edge detection. Mapping of these features is rendered and compared directly with the respective view. Discrepancy between a 3D hand pose and the actual observation is quantified by an error function minimized through particle swarm optimization. The pose for which this error

function is minimal constitutes the output of the proposed method at a given moment in time. As a temporal continuity in hand motion is assumed, initial hypotheses for current time instance are restricted in the vicinity of the previous time instant solution. Being computationally expensive, the method is implemented in a GPU, resulting in near real-time performance. Their study is improved in [15] using Kinect® and a single hypothesis, where observation is the RGB-D image segmented to locate the hand through skin color and depth of the scene; therefore, error function is different, computational requirements are lower, camera array is simplified, and resulting system works even under variable lighting conditions.

On systems that do not handle occlusions or interactions with other objects, certainty in estimating hand positions is seriously affected, so the role of context in object recognition is very significant [16]. Several researchers have tried to exploit the contextual constraints on Computer Vision problems, in [17] a brief count of researching work considering the context in the classification of human-object interaction activities can be found, differentiating between those who have focused on the human body or hands and those who provide a detailed 3D model of them and the object. This project is an extension of that presented in [11] by considering jointly the hand and the manipulated object. It is an optimization problem whose solution is the 26 DOF hand pose along with the pose and parameters of the manipulated object model using a multi-camera system. In each of the acquired observations, skin color maps and edges of the hand are extracted; depending on the point of view, the presence of an object can obstruct the presence of the hand, their incomplete observation provides evidence of the type and pose of the manipulated object and at the same time the object improves the estimation of hand pose. The process seeks the hand-object model that best explains the incompleteness of the resulting observations of the occlusions derived from their interaction and also be physically plausible (that the hand does not share the same physical space with the object) by penalty the objective function. Regarding methodology, the authors use Canny edge detection to build an edge map, compute a distance transform for each one, and use a previous own method to generate the color map. Thus, the image observations are given by the skin color maps and the transform. The authors claim that this is the first model-based work that efficiently solves the continuous full-DOF, joint hand-object tracking problem based solely on markerless multi-camera input, further demonstrating that hand-object interaction can be seen as a context that facilitates hand pose estimation, instead of being a problem factor.

Ren et al [18] propose a distance metric called "Finger-earth mover's distance" to measure the dissimilarity of the noisy hand shape provided by a Kinect® sensor, as method just matches fingers and not the whole hand shape, it can better distinguish hand gesture subtle movements. This metric sees each finger as a "cluster", penalizing unmatched fingers. The method is proposed to address the problem that, due to the low resolution of the depth map delivered by the Kinect® sensor (640x480), it is hard to detect and segment a small object

like the hand and all its joints.

Oikonomidis et al [19] extend again their work with a model-based, joint-evidence method, where a two-hand tracking is performed as an optimization problem whose objective function quantifies the discrepancy between the structure and 3D appearance of hypothetical configurations of both hands and the corresponding Kinect® observations. Optimization is performed by a variant of a particle swarm optimization method, adapted to the needs of the specific problem. The methodology combines the steps performed in their previous studies [15] and [17], especially in the latter idea to model the hand-object relations and to treat occlusions as a source of information rather than see them as a complicating factor. Furthermore, in this work the problem is more complex since it focuses on both hands with only one Kinect® sensor instead of a multi-camera system. An update of this work can be found likewise in [7].

In [20], a method to capture the articulated motion of two hands while interacting with each other and with an object is proposed. Salient points such as finger tips are scanned through a multi-camera system, however, since these points cannot be tracked continuously due to excessive occlusions and similarity in their features and color appearance, avoiding a fixed association between the salient points and the respective fingers, an approach that solves the salient point association jointly with the hand pose estimation problem is proposed. Also, a quite differentiable objective function for pose estimation is implemented, taking into account edges, optical flow, salient points, and collisions. Thus, authors may use simple local optimization instead of a sampling based one as in [19]; in fact, they say their approach achieves significantly lower pose estimation errors than the sampling optimization. In conclusion, they suggest the possible desirability of researching the combination of both optimization techniques.

In [2], a new approach for tracking 3D articulated skeletal models using an augmented rigid body simulation is presented, being able to follow a human hand from a depth sensor. The method allows robust, real-time results using only an x86 processor. The system generates constraints that limit motion orthogonal to the rigid body model's surface, these constraints, along with prior motion, collision constraints, and joint mechanics, are solved by a Gauss-Seidel solver. To improve tracking accuracy, multiple simulations are generated at each frame and fed some heuristics, constraints, and poses.

Kulshreshth et al [21] present preliminary results of a real-time, markerless finger tracking technique using a Kinect® sensor as an input device. The technique calculates feature vectors based on Fourier descriptors of equidistant points chosen on the silhouette of the detected hand and matches templates to find the best fit.

Karnan et al [22] propose a method to control the movement of a mouse pointer using simple hand gesture 2D images and a webcam. An algorithm for real-time tracking based on adaptive skin detection and motion analysis is implemented. Using the history of motion, the trajectory of movement of the hand is drawn and used to identify a gesture. The image database consists of four different gestures. In order to scale the motion when user is far away from

the point of capture, an algorithm is used to define the region of interest, motion of the mouse pointer is scaled accordingly. The system is fully automatic, real time, and does not need a uniform background.

In [23] a method for real-time continuous pose recovery of markerless complex articulated objects from a single depth image is described. In order to generate the training data, the system can use multiple depth cameras, however, only a single depth camera for real-time tracking is required. The method can be generalized to track any articulated object that (a) can be modeled as a 3D boned mesh, (b) can be fed to a binary classifier to label pixels belonging to the object, and (c) that the projection from bones pose space to a 2D depth image be approximately one to one. Four stages are distinguished:

1. a randomized decision forest classifier for image segmentation,
2. a robust method for labeled dataset generation,
3. a convolutional network for dense feature extraction, and
4. an inverse kinematics stage for stable real-time pose recovery.

6 Main Expected Contribution

Oikonomidis et al [11] suggest that there is great interest in the development of markerless, computer vision based solutions, since they are not invasive and maybe less expensive. Furthermore, by fully understand hands configuration thanks to their 3D pose estimation, systems that understand human activities and interaction with their physical and social environment could be built. The economic benefit that areas such as ludic get globally, and all the advantages that could bring the development of these new ways of communication to the daily life of every human being, encourage scientific community to further research and improve or develop new methods looking for a higher efficiency and accuracy. But above all, this study was conducted to provide background on the research area as a basis for developing a set of tools that can be applied in the handling of HCI devices by people with motor disabilities, whose condition has not been actually addressed by the current hand gesture recognition methods.

7 Discussion of the Results and their Validity

In this paper, a brief review of recent research efforts in hand gesture recognition has been performed. Table 1 is a comparative summary of the tools, features, techniques, and objectives reviewed. As shown, several areas of opportunity can be derived from these data, regarding current research in the literature. The following are of particular interest for the purpose of this study, therefore will be addressed in the development of the project.

1. The use of two or maybe more RGB-D cameras (whether Kinect® or other brands), and/or other technology such as optical flow or infrared light sensors, could mean a significant advantage mainly to avoid occlusions in scanned objects.

2. A combination of techniques concerning feature extraction and optimization methodology to check if there is an improvement (or optimization) on recognition.
3. The application of these approaches to people with motor or speech disabilities, which has not been addressed in the state-of-the-art, and besides being a relevant research topic, becomes an increasing needing for them to interact with various technological devices.

Table 1. Abbr: DT=Distance Transform, PSO=Particle Swarm Optimization, AD=Adaptive detection, RDFC=Randomized Decision Forest Classifier, CN=Convolution Network, IK=Inverse Kinematics, HFC=Hough Forest Classifier, FEMD=Finger-Earth Mover's Distance, FD=Fourier Descriptors

Research	Camera	Features	Technique	Optimization	Objective
[11]	Multiple	Skin color, Edges	DT, Canny	PSO	One hand
[15]	Kinect®	Skin color, Depth	-	PSO	One hand
[22]	Webcam	Skin color	AD	-	One hand
[23]	Depth	Depth	RDFC, CN	IK	One hand
[2]	Depth (2 sensors)	Depth	-	Gauss-Seidel	One hand to two hand
[17]	Multiple	Skin color, Edges	DT, Canny	PSO	One hand-object interaction
[19]	Kinect®	Skin color, Depth	-	PSO	Two hand interaction
[20]	Multiple	Edges, Optical flow, Collisions, Salient points	HFC	Local	Two hand-object interaction
[18]	Kinect®	Skin color, Depth	FEMD	-	Fingers
[21]	Kinect®	Depth, Silhouette	FD	-	Fingers

References

1. Nakashika, T., Hori, T., Takiguchi, T.: Depth Spatial Pyramid: A pooling method for 3D-object recognition. Advances in Computer Science and Engineering 12, 15–30 (2014)
2. Melax, S., Keselman, L., Orsten, S.: Dynamics based 3D skeletal hand tracking. In: Proceedings of Graphics Interface 2013, pp. 63–70 (2013)
3. Bachmann, D., Weichert, F., Rinkenauer, G.: Evaluation of the Leap Motion Controller as a new contact-free pointing device. Sensors 15(1), 214–233 (2014)
4. Kim, D., Hilliges, O., Izadi, S., Butler, A.D., Chen, J., Oikonomidis, I., Olivier, P.: Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In: proceedings of the 25th annual ACM symposium on User interface software and technology, pp. 167–176 (2012)

5. Jung, P., Lim, G., Kim, S., Kong, K.: A wearable gesture recognition device for detecting muscular activities based on air-pressure sensors. IEEE Transactions on Industrial Informatics 11, 485–494 (2015)
6. Wang, R.Y., Popović, J.: Real-time hand-tracking with a color glove. ACM Transactions on Graphics 28(9), 63:1–63:8 (2009)
7. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of human hands in 3D, ERCIM News 95, pp. 23–25 (2013)
8. Zimmerman, T.G., Lanier, J., Blanchard, C., Bryson, S., Harvill, Y.: A hand gesture interface device. In: Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface, pp. 189-192. ACM Press, New York, USA (1987)
9. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding 108(1), 52–73 (2007)
10. Salzmann, M., Urtasun, R.: Combining discriminative and generative methods for 3D deformable surface and articulated pose reconstruction.In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 647–654 (2010)
11. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Markerless and efficient 26-dof hand pose recovery. In: Kimmel, R., Klette, R. and Sugimoto A. (eds.), ACCV 2010, Part III. LNCS, vol. 6494, pp. 744–757. Springer Verlag Heidelberg (2011)
12. Bray, M., Koller-Meier, E., Van Gool, L.: Smart particle filtering for high-dimensional tracking. Computer Vision and Image Understanding 106(1), 116–129 (2007)
13. Heap, T., Hogg, D.: Towards 3D hand tracking using a deformable model. In: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, pp. 140–145 (1996)
14. Lin, J.Y., Huang, T.S.: Capturing natural hand articulation. In: Proceedings of the Eighth IEEE International Conference on Computer Vision, pp. 426–432 (2001)
15. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3D tracking of hand articulations using Kinect. In: procedings of British Machie Vision Conference, pp. 1–11 (2011)
16. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in cognitive sciences 11(12), 520–527 (2007)
17. Oikonomidis, I., Kyriazis, N., Argyros, A. a.: Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: International Conference on Computer Vision, pp. 2088–2095 (2011)
18. Ren, Z., Yuan, J., Zhang, Z.: Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In: proceedings of the 19th ACM international conference on Multimedia, pp. 1093–1096 (2011)
19. Oikonomidis, I., Kyriazis, N., Argyros, A. A.: Tracking the articulated motion of two strongly interacting hands. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1862–1869 (2012)
20. Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: Fitzgibbon et al. (eds.), Part VI. LNCS, vol. 7577, pp. 640–653. Springer Verlag Heidelberg (2012)
21. Kulshreshth, A., Zorn, C., LaViola, J.J.: Poster: Real-time markerless kinect based finger tracking and hand gesture recognition for HCI. In: IEEE Symposium on 3D User Interfaces, pp. 187–188 (2013)
22. Karnan, J., Ramkumar, M., Sivaraman, K., Santhakumar, G., Karthik Kumar, R.: Real-time gesture based human computer interaction for office applications.

International Journal of Review in Electronics and Communication Engineering 2(1), 42–47 (2014)
23. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics 33(5), 69:1–69:10 (2014)