Recognizing Intuitive Pre-defined Gestures for Cultural Specific Interactions: an Image-based Approach

Marco Roccetti
Computer Science Department
University of Bologna
Bologna, Italy
roccetti@cs.unibo.it

Gustavo Marfia
Computer Science Department
University of Bologna
Bologna, Italy
marfia@cs.unibo.it

Abstract— **Hands-free gaming systems are likely to gain, in the near future, the total share of the electronic gaming market. Such process has been anticipated by the introduction of the Nintendo Wii that, within a few months from its launch, stably reached the top ranking of sales among gaming systems. We can expect something similar will occur once the for-long expected Microsoft Kinect system will be available on the shelves of computer gaming stores. In general, it is clear from this sequence of events, that the opportunity of playing a game naturally, performing the same movements that would be made in a real setting, has a magnetic effect on customers. What, instead, is less evident is the great wealth of research that has been performed in the recent years to reach such result. In fact, behind the development of such gaming platforms strenuous research efforts have been made in a number of areas, which include and are not limited to hand following techniques and gesture recognition algorithms. As a beneficial side effect, the development of these areas can also play an important role in a number of other applications, including technical and artistic interactive exhibitions. In this paper we will describe the process that led us to devise a novel hand following and an innovative gesture recognition technique that both can easily be applied to exhibition scenarios (e.g., museums, fairs, etc.) and in general all those performing events where a pre-defined set of gestures are needed to enjoy a cultural experience, as they can be easily jointly put to good use to recognize a predefined set of movements. Our algorithms are robust and easy to implement, for this reason they are particularly suited for exhibition scenarios where stands can often change or adapted to new requirements.**

Hand follower; gesture recognition; technical exhibition; artistic exhibition.

I. INTRODUCTION

Gaming platforms such as the Nintendo Wii and possibly greatly expected new ones, such as the Microsoft Kinect and Sony Move, have and will radically change how gamers interact with computers. Clearly, their important merit has been that of making digital technological advancements, such as hand and gesture recognition algorithms, easily appreciable by all. In fact, the use of such systems has generally raised the level of the expectations for a whole set of entertainment applications, other than computer games. An example of such phenomenon, that has progressively increased its importance in the past few years, has been a prolific use and application of a wealth of different technological tools in museums, entertainment parks, technical fairs, etc.

Among other possible locations designed for entertainment purposes, technical and artistic exhibitions, especially, have progressively become more sensitive to the use of new technological tools and gadgets. In fact, visitors often expect that organizers provide more effective and impressive ways of entertaining and educating, not limiting their visiting experience to a simple observation of a set of objects and artifacts. Examples that have worked on meeting such demand are those museums that installed haptic interfaces capable of giving the impression of virtually touching the pieces of art that they preserve [1], and the cathedrals and monuments that show, through special LCD screens, the exact reconstruction of how their surrounding scenery would have looked like centuries ago [2], [3].

However, thanks to its most recent advancements, technology has found even more ways of integration with technical and artistic exhibitions [8-10]. In fact, the most recent technological findings made in fields that include hand and gesture recognition algorithms have paved the way to performing exhibitions where a visitor, when observing something of interest shown at a stand or on a stage, can also interact with it, appreciating specific effects which correspond to a given set of movements and gestures. Such scenario, clearly, opens the gates to a whole set of applications where, interestingly, it is possible to learn and enjoy an experience by not only watching and listening, but also by dynamically interacting with an exhibition.

The main scope of this paper is to contribute to the interactive process that can be established between an exhibition and its visitors, using modern digital communication technologies. To pursue this task, we devised a system that can follow and recognize the gestures a visitor performs. Compared to solutions alike in scope based on modification of the Nintendo Wii, ours does not require that visitors wear any sensor or gadget. Moreover, our solution does not locate a visitor's hands utilizing estimation schemes based on the

color tone of the skin, thus it is robust to changes in color and to differences in illumination exposure.

In brief, our solution is based on the idea that visitors may perform movements within a set of pre-defined gestures. The way through which visitors learn admissible gestures exploits the use of an LCD monitor and a video camera. A visitor at first reads or watches a demonstration of the pre-defined movements it should mimic on an LCD monitor that is positioned in front of him or her. Once this phase has terminated, the video camera enters the scene. In fact, a video camera is located above the visitors and is used to record the movements they perform, feeding such information to a software module that implements our hand following and recognition system. Once the hands of a visitor are tracked, they are digitally represented on the screen, thus each movement performed in reality is followed and echoed on the screen. At this point, the visitor can perform a set of gestures, interacting with the digital world shown on the screen. Based on the resemblance of the visitor's movements with any of the pre-defined ones that are stored in the system, the system reacts in a different way. Therefore, depending on which is the purpose of the exhibition, it is possible to give different forms of feedback corresponding to the different type of movements a visitor may perform.

The aforementioned functionalities, from a software point of view, are implemented using three main modules: a hand follower, a gesture recognizer and an audiovisual feedback displayer. In this paper, we focus on the first two modules, describing how we devised a set of algorithms that can simply analyze a video source, locating a person's hands position and recognizing the gestures it performs. Moreover, in order to provide a fair comparison that may better exemplify the advantages introduced by our system, its robustness and efficiency in recognizing a visitor's movements, we implemented two additional hand follower systems based on alternative approaches, the Wii remote and the OpenCV libraries [4], [5]. The rationale behind this choice was to compare our method with possible alternatives in order to assess its efficacy. All these methods have then been tested on a case scenario we here briefly describe in the following, where an exhibition system teaches cooks how to prepare filled fresh pasta from raw ingredients, to be exploited in a culinary fair.

The rest of this paper is organized as follows. In the next Sections II and III we will directly move to the core of our algorithms, explaining how they have been devised and we will describe their performances. We will then detail how the algorithms are integrated on a real testbed in Section IV, in order to be deployed on a scenario of interest. We finally conclude with Section V.

II. HAND FOLLOWER MODULE

The hand follower is, clearly, one of the most complex modules as, in a realistic setting, many possible sources of noise could emerge (e.g., different hand sizes, different skin color tones, etc.). In the following, we will begin describing our solution and then contrast it with two alternative solutions, based on the Wii remote and on the used of the OpenCV libraries.

In brief, our new solution is based on storing, at the very beginning, a frame that represents the background image taken from the above standing webcam. After this first step is performed, our algorithm keeps comparing the current image frame to the background image previously stored. If nothing is in sight, the difference between the two images is a null image that indicates that nothing entered the observed area. If, instead, something or someone enters the area, the difference image possibly represents the hands of a visitor. Clearly, such solution would result to be too simple if any part of the human body could enter the observed area. In our setting, however, we expect that a visitor will only be able to stretch his hands and arms over the area that is observed by the webcam, so any positive match will be necessarily caused by such parts of the human body. This algorithm, as we have now described, is outlined in Figure 1 and does not require any further explanation. However, also such solution comes with evident advantages and disadvantages. Its main advantage is its high performance in terms of speed, as the functionalities it implements are very simple and do not go beyond simple difference operations between given frames. Its disadvantages, on the other hand, are caused by its simplicity, as it is not really able to distinguish the parts of the human body, since every given object moving in front of the video camera can generate a false positive.

Now, before entering the details of how a solutions based on the Wii remote works, we should mention that its use has been thoroughly investigated in the past few years, as recent research has shown that it can be successfully turned into a motion-tracking device [4]. In practice, the Wii remote solution is stable and mature, as its software libraries have been used for long now and the accuracy of its sensors is sufficient for our purposes. However, we should remind that using it in an exhibition scenario entails forcing any visitor to wear an infrared light source. Clearly, in a context where many thousand visitors may daily interact with a system (e.g., the Vatican museums in Rome), such solution may result unfeasible. For this reason, thanks to its performance the Wii remote solution will serve as our performance benchmark, although it does not represent our target solution.

As a second alternative possibility, we built a solution based on the OpenCV libraries. Differently from the Wii remote solution, where each visitor should have worn an infrared light emitting source, the OpenCV libraries have been designed to analyze, extract and recognize given shapes and patterns from an image or a video stream captured from a webcam. Among the many functionalities supported by such libraries, they also provide a set of functions that can be used to identify the areas of a given image that are closest to a given color. In brief, considering that the color tone of the skin of a person is typically not uniform, the best way to individuate a person's hands is to select the two areas of a given minimum size where the color tone lies within two thresholds. Once these areas have been identified, it is possible to compute the position of the center of mass of the two hands averaging the

1: *background* ← background color acquisition.
2: *Left(t, background)* ← difference pixels in the left hand area at time t;
3: *Right(t, background)* ← difference pixels in the right hand area at time t;
4: *leftHandY(t)* ← *y* of any pixel in *Left(t, background)*;
5: *leftHandX(t)* ← *x* of any pixel in *Left(t, background)*;
6: *rightHandY(t)* ← *y* of any pixel in *Right(t, background)*;
7: *rightHandX(t)* ← *x* of any pixel in *Right(t, background)*;

Figure 1. New hand follower.

planar coordinates of the contours that embrace the two areas. Interestingly, differently from the Wii remote case, using the OpenCV libraries it is possible to identify the center of mass of the two hands of a person without the need of using any additional hardware, but a video camera. However, using such technique comes with some problems as well. The first problem we recognized is related to the identification of the center of mass of the two hands. This process, in fact, requires finding which are the areas occupied by the two hands. However, also the arms of a person, can, for example, interfere. For this reason, using a skin detection algorithm alone can bring to an incorrect estimation of the location where the hands of a visitor are, as also other parts of its body could be mistakenly processed. We dealt with this problem devising another algorithm, that, instead of using a technique that computes the center of mass of the areas occupied by a person's skin, simply finds the position of the two hands as the two farthest ahead points occupied by the detected areas. In fact, since the video camera is placed above the head of a user and we expect that any hand gesture will be made in front of a visitor, we found this to be an effective way to practically pinpoint the two hands. This second algorithm is depicted in Figure 2, where all the steps pertaining the identification of the two hands of a visitor are shown. In particular, at first the algorithm learns the color tone range of a given visitor (line 1). After this first step has been performed, it identifies the two areas where the hands of a visitor lie (lines 2 and 3). The final step, performed in lines 4 through 7, is then to select the farthest ahead point of the two areas.

Clearly, at this point, there seems to be no evident winner in the match between the OpenCV-based approach and our new one, as both could be applied within our case scenario. For this reason, the best way to understand which methodology was best the best match for our application scenario was implementing and testing both systems. Both, in fact, showed good results in tracking the hands of a visitor, producing a responsiveness delay (i.e., the time lag between the time a movement occurs and the time when the movement can be appreciated on the screen) that was always confined within ten milliseconds. Performances disagreed, instead, when we compared the two methodologies in terms of the frequencies with which they were able to correctly identify the positions of the two hands. In particular, when we used a cheap webcam, the OpendCV-based approach correctly identified the hands of a person with a higher rate (almost 100% hit rate), compared to our new one (around 80% hit rate). These results, however, inverted when we used an advanced IP camera with a higher resolution and refresh rate. The explanation of this phenomenon may be understood considering that our webcam was provided with a self-stabilizing system that automatically filtered out all the shadowing effect, while the advanced professional camera required to be manually tuned every time, for each different setting. Clearly, when the illumination effects were not filtered out, the OpenCV-based algorithm had a hard time recognizing a visitor's skin, as the color tone could greatly differ from the one that was previously acquired by the system.

III. GESTURE RECOGNIZER MODULE

Once a visitor's hands can be identified, the challenge becomes that of recognizing when a given gesture is performed. In order to implement an efficient and robust gesture recognizer, we devised a method that is simply based on geometrical considerations. In fact, any gesture may be simply represented identifying a starting point, an ending point and a trajectory. As we will briefly see, such approach has validly served our purpose.

The majority of gestures, except those that include some periodic circular movement, may be schematically represented using two areas and a trajectory that connects the two areas. Moreover, many gestures are characterized by not only their geometric characteristics, but also by the time that is employed to complete them. As an example, consider a game where a player is boxing against an avatar, in such scenario faster movements should correspond to harder punches. For this reason we also associated a timer to each movement, which gives the possibility to our algorithm to verify whether the given gesture has been performed within its pre-defined time constraints. This timer also has a second scope, as we should also consider the case where a visitor decides to abort its interactive experience, leaving the exhibition stand. In this way, if a timer triggers the system to restart once certain amount of time has elapsed without anything happening, the system will not wait indefinitely for someone to move and some other visitor will be given the chance of interacting with the computer exhibition.

The abovementioned considerations have all been implemented in Figure 3. In particular, we can see that at line 1 our algorithm tests whether a visitor's hand has been located within the initial area. In the positive case, a timer is started (line 2). Until the visitor does not move his hand within the final area (line 4), the gesture recognition module keeps checking whether two events have happened (line 5): a) the visitor's hand left the correct trajectory and b) the timeout fired. If before any of the two above events occur the

```
1: learnColorRange ← learn visitor's skin color tone.
2: Left(t, colorRange) ← Skin pixels in the left hand area at time t;
3: Right(t, colorRange) ← Skin pixels in the right hand area at time t;
4: leftHandY(t) ← y s.t. y = max Left(t, colorRange);
5: leftHandX(t) ← x s.t. y = max Left(t, colorRange);
6: rightHandY(t) ←y s.t. y = max Right(t, colorRange);
7: rightHandX(t) ← x s.t. y = max Right(t, colorRange);
```

Figure 2. OpenCV-based hand follower.

```
1:  check (handPosition == initialArea) {
2:      handPosition ← getCurrentPosition();
3:      timerIsStarted();
4:      while (handPosition != finalArea) {
5:          if (handPosition != trajectory or timeoutFired) {
6:              mistake();
7:          }
8:      }
9:      ok();
10: }
```

Figure 3. Gesture Recognizer.

visitor's hand lands on the final area, the function leaves successfully recognizing that a correct gesture has been performed (line 9).

IV. Preparing Fresh Pasta

To witness that our system works to recognize a pre-defined set of gestures, we apply it to a culinary case. Now, we can explain how our system can be put to good use in an interactive exhibition system. For this purpose, we designed a system that teaches how fresh pasta can be prepared starting from its raw ingredients: eggs, flour and water. Clearly, such process is composed of various steps, each of which is paramount in reaching the desired final result. Therefore, our first task has been that of identifying those steps where a visitor could easily watch and perform a set of gestures. As an example, consider the case where the eggs are laying on the side, while the flour has already been poured on the table (Figure 4). At first a digital recording explains the visitor that it should grab the eggs, break them and pour their content on the flour. This summarizes one step, where we can easily find a starting position (the position where the eggs lie), an ending position (the position where the flour lies), and a trajectory (the shortest path between the eggs and the flour).

How each step has been organized is briefly represented in Figure 5. At first a visitor watches what kind of gestures it should perform and then accomplishes the required movements. In the meantime, its hands are echoed on the display, its movements are tracked by our hand follower and checked by our gesture recognition system for their correctness. In case the set of movements that were initially described were correctly mimicked, their result would be shown on the screen. As an example, consider Figure 6, which results from a visitor performing the required step when beginning from the situation shown in Figure 4.

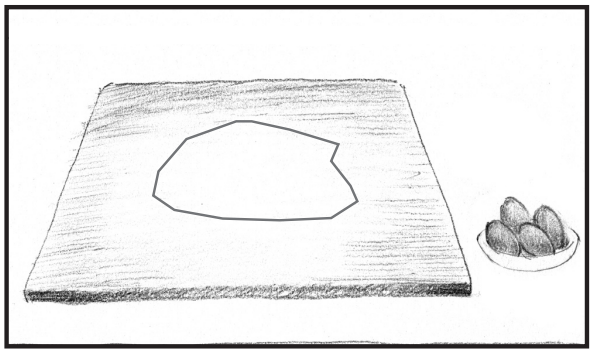

Figure 4. Table board situation before the eggs are broken on the flour.

Interested readers can refer for more details to [6], where we thoroughly explain how our system has been implemented, giving a deeper understanding of its scope and results.

V. Conclusion

We described, in this paper, the process that led us to devise a novel hand following and an innovative gesture recognition technique that can jointly be applied to exhibition scenarios (e.g., museums, fairs, etc.). Contrary to many other applications of gesture following/recognizing

techniques or similar schemes [7-21], applying gesture recognition in a cultural/exhibition context has required the devise of innovative algorithms optimized for this specific field. We have described our gesture recognizer technique which, based on geometric considerations, can easily detect when a pre-defined gesture has been completed. Our algorithms are easy to implement and particularly robust and, as such, suited for exhibition environments subject to high volumes of daily visitors.

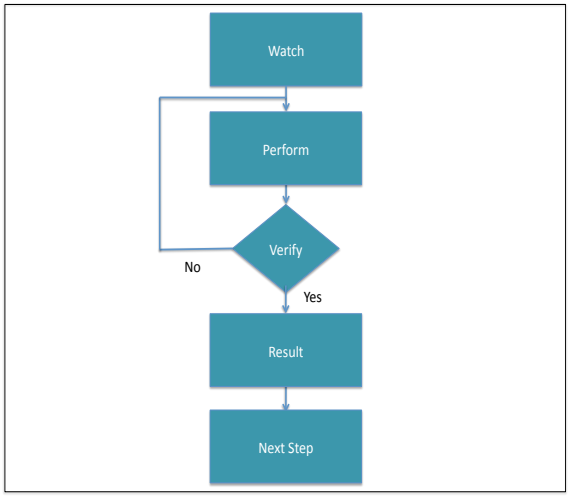

Figure 5. How each step of our system can be represented.

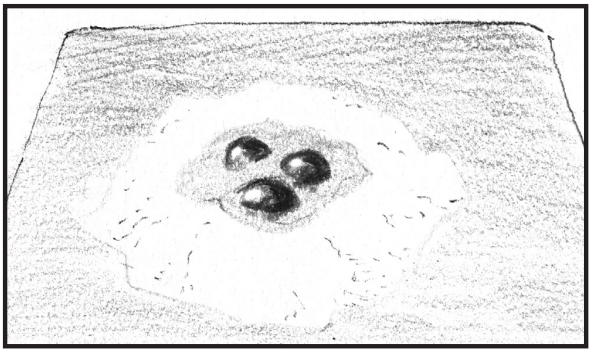

Figure 6. Table board situation right after the eggs have been broken on the flour.

REFERENCES

[1] M. Carrozzino, C. Evangelista, A. Scucces, F. Tecchia, G. Tennirelli and M. Bergamasco, "The virtual museum of sculpture", in Proceedings of the 3rd ACM International Conference on Digital interactive Media in Entertainment and Arts, 2008, Athens, Greece, pp. 100–106.

[2] S. El-Hakim, J. A. Beraldin, M. Picard, "Detailed 3D reconstruction of monuments using multiple techniques", in Proceedings of ISPRS-CIPA Workshop, Corfu, Greece, 2002, pp. 13–18.

[3] B. Frischer, D. Abernathy, G. Guidi, J. Myers, C. Thibodeau, A. Salvemini, P. Muller, P. Hofstee, B. Minor, "Rome reborn", ACM SIGGRAPH 2008 new tech demos, 2008.

[4] J. C. Lee, "Hacking the Nintendo Wii Remote", IEEE Pervasive Computing, July–September 2008, pp. 39–45.

[5] S. Ferretti, M. Roccetti, and F. Strozzi, "On developing tangible interfaces for video streaming control: a real case study," in Proceedings of 18th ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video, Braunschweig, Germany, pp. 51-56.

[6] M. Roccetti, G. Marfia, M. Zanichelli, "The art and craft of making the tortellino: playing with a digital gesture recognizer for preparing pasta culinary recipes", University of Bologna Technical Report, 2010.

[7] A. Semeraro, "Minority report: a case study", the Made in UNIBO Exhibition, Museum of Modern Art of Bologna, 2008.

[8] F. Bevilacqua, B. Zamborlin, A. Sypniewsky, N. Schnell, F. Guedy, N. Rasamimanana, "Continuous realtime gesture following and recognition", Embodied Communication and Human-Computer Interaction, Lecture Notes in Artificial Intelligence n. 5934, Springer, 2010, pp. 73-84.

[9] C.E. Palazzi, M. Roccetti, "A Groupware for Pigment Identification in Cultural Heritage", International Journal of Virtual Reality, IPI Press , vol. 8, no. 3, 2009, pp. 51-56.

[10] A. Hamidian, C.E. Palazzi, T.Y. Chong, M. Gerla, U. Körner, "Exploring Wireless Mesh Networks for Collaborative Augmented Reality Environments", International Journal on Virtual Reality, IPI Press, vol. 9, no. 2, Jun 2010.

[11] K. Suzuki, S. Hashimoto, "Robotics Interface for Embodied Interaction via Dance and Musical Performance", in Proceedings of IEEE, vol. 92, no. 4, 2004, pp. 656-671.

[12] N. Tosa, "Expression of Emotion Unconsciousness with Art and Technology", Affective Minds, G. Hatano, N. Okada, and H. Tanabe, eds., Elsevier, 2000, pp. 183-201.

[13] C. Pinhanez and A. Bobick, "It/I: A Theater Play Featuring an Autonomous Computer Character," Presence: Teleoperators and Virtual Environments, vol. 1, no. 5, 2002, pp. 536-548.

[14] A. Bierbaum, C. Just, P. Hartling, K. Meinert, A. Baker, C. Cruz-Neira, "VR Juggler: A Virtual Platform for Virtual Reality Application Development", in Proceedings of IEEE VR 2001, Yokohama, Japan, 2001.

[15] M. Bull, P. Gilroy, D. Howes, D. Kahn, "Introducing Sensory Studies", The Senses and Society, March 2006, pp. 5-7.

[16] M. Weal, D. Cruickshank, D. Michaelides, D. Millard, D. De Roure, E. Hornecker, J. Halloran, G. Fitzpatrick, "A reusable, extensible infrastructure for augmented field trips", in Proceedings of 2nd IEEE International Workshop on Pervasive ELearning, pp. 201–205.

[17] M. Roussou, "Learning by Doing and Learning Through Play: an exploration of interactivity in virtual environments for children", in Computers in Entertainment, vol. 2, no. 1, section: Virtual reality and interactive theaters, ACM Press.

[18] Y. Rogers, M. Scaife, S. Gabrielli, H. Smith, E. Harris, "A Conceptual Framework for Mixed Reality Environments: Designing Novel Learning Activities for Young Children", in Presence Teleoperators and Virtual Environments. Vol. 11, no. 6, pp. 677–686.

[19] S. Price, Y. Rogers, "Let's get physical: the learning benefits of interacting in digitally augmented physical spaces", Computers & Education, Vol. 43, pp. 137–151.

[20] N. Pares, A. Carreras, J. Durany, "Generating meaning through interaction in a refreshing interactive water installation for children", in Proceedings of Interaction Design and Children 2005, ACM Press.

[21] A. Johnson, M. Roussos, J. Leigh, C. Vasilakis, C. Barnes, T. Moher, "The NICE project: learning together in a virtual world", in Proceedings of the IEEE Virtual Reality Annual International Symposium.