Navigating Wall-sized Displays with the Gaze: a Proposal for Cultural Heritage

Davide Maria Calandra, Dario Di Mauro, Francesco Cutugno, Sergio Di Martino
Department of Electrical Engineering and Information Technology
University of Naples "Federico II"
80127, Naples, Italy
{davidemaria.calandra, dario.dimauro, cutugno, sergio.dimartino}@unina.it

ABSTRACT

New technologies for innovative interactive experience represent a powerful medium to deliver cultural heritage content to a wider range of users. Among them, Natural User Interfaces (NUI), i.e. non-intrusive technologies not requiring to the user to wear devices nor use external hardware (e.g. keys or trackballs), are considered a promising way to broader the audience of specific cultural heritage domains, like the navigation/interaction with digital artworks presented on wall-sized displays.

Starting from a collaboration with a worldwide famous Italian designer, we defined a NUI to explore 360°panoramic artworks presented on wall-sized displays, like virtual reconstruction of ancient cultural sites, or rendering of imaginary places. Specifically, we let the user to "move the head" as way of natural interaction to explore and navigate through these large digital artworks. To this aim, we developed a system including a remote head pose estimator to catch movements of users standing in front of the wall-sized display: starting from a central comfort zone, as users move their head in any direction, the virtual camera rotates accordingly. With NUIs, it is difficult to get feedbacks from the users about the interest for the point of the artwork he/she is looking at. To solve this issue, we complemented the gaze estimator with a preliminary emotional analysis solution, able to implicitly infer the interest of the user for the shown content from his/her pupil size.

A sample of 150 subjects was invited to experience the proposed interface at an International Design Week. Preliminary results show that the most of the subjects were able to properly interact with the system from the very first use, and that the emotional module is an interesting solution, even if further work must be devoted to address specific situations.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Interaction styles

1. INTRODUCTION

Wall-sized displays represent a viable and common way to present digital content on large projection surfaces. They are applied in many contexts, like advertisement, medical diagnosis, Business Intelligence, etc. Also in the Cultural Heritage field, this type of displays is highly appreciated, since they turn out to be particularly suited to show to visitors artworks that are difficult or impossible to move, being a way to explore the digital counterpart of real/virtual environments. On the other hand, the problem with these display is how to mediate the interaction with the user. Many solutions have been proposed, with different trade-off among intrusiveness, calibration and precision degree to be achieved. Recently, some proposals have been developed aimed at exploiting the direction of the gaze of the visitor in front of the display as a medium to interact with the system. The simple assumption is that whether the user looks towards an edge of the screen, he/she is interested in discovering more content in that direction, and the digital scenario should be updated accordingly. In this way, there is no need to wear a device, making easier for a heterogeneous public to enjoy the digital content.

Detecting the gaze is anyhow a challenging task, still with some open issues. To estimate the Point of Gaze (PoG), it is possible to exploit the eye movements, the head pose or both [23], and to require special hardware to wear (e.g.: [12]) or to develop remote trackers (e.g.: [6]). The latter are not able to provide a high accuracy, but this is an acceptable compromise in many scenarios, like the Cultural Heritage, where the use of special hardware for the visitors is usually difficult.

For the Tianjin International Design Week 2015[1], we were asked to develop a set of technological solutions to improve the fruition of a 360°digital reconstruction projected on a wall-sized display of the "*Camparitivo in Triennale*"[2], a lounge bar (see Figure 1) located in Milan, Italy, designed by one of the most famous Italian designers, Matteo Ragni, to celebrate the Italian liqueur *Campari*. The requirements for the solution were to define a Natural User Interface (NUI), which does not constrain users to maintain a fixed distance from the display, neither to wear an external device.

To achieve our task, we designed a remote PoG estimator for wall-sized displays where 360°virtual environments are rendered. A further novelty element of the proposal is the exploitation of another implicit communication channel of the visitor, i.e. his/her attention towards the represented image on the display. To this aim, we remotely monitor pupil size variations, as they are significantly correlated with the arousal level of users while performing a task. This information can be firstly useful to the artist, as pupils dilate when visitors are looking at pleasant images [9]. Moreover, logging the pupil dilation (*mydriasis*) during an interaction session can be a reliable source of information, useful also to analyze the usability

DOI:

[1]http://tianjindesignweek.com/
[2]http://www.matteoragni.com/project/camparitivo-in-triennale/

Figure 1: Matteo Ragni's "Camparitivo in Triennale"

level of the interface, since pupils dilate when users are required to perform difficult tasks, too [11] [3].

In this paper we describe both the navigation with the remote PoG estimator and the solution for logging the mydriasis, together with a preliminary case study. More in details, the rest of the paper is structured as follows: in section 2, we explain the navigation paradigm for cultural content with the gaze, detailing the steps we performed to detect and track the PoG. In section 3, we explain how the mydriasis detection could be a useful strategy to investigate the emotional reactions of users enjoying a cultural content and we detail our steps to get the pupil dilation. In section 4, we present the case study: Matteo Ragni's *Camparitivo in Triennale*, showing how we allow the visitors to navigate the digital rendering of the lounge bar, on a wall-sized display, reporting some preliminary usability results. Section 5 concludes the paper, presenting also future research directions.

2. NAVIGATING WITH THE GAZE

Even if wearable eye trackers are becoming smaller and more comfortable, they still have an impact on the quality of a cultural visit. We believe that the user experience strongly depends on the capability of the user to establish a direct connection with the artworks, without the mediation of a device. For this reason, in order to allow the user to explore a 360° cultural heritage environment using only his/her point of gaze, we focused on developing a remote head pose estimator for wall-sized displays, which does not require users to wear any external device or to execute any prior calibration.

The contents that we aim to navigate are 360°virtual environments, expressed as a sequence of 360 frames whose step size is 1°. Thus, navigating the content on the left (right) means to show the previous (next) frame of the sequence. As we want visitors to feel the sensation of enjoying an authentic large environment, the wall-sized display is used to represent the content with real proportions. If by one side, this choice improves the quality of the fruition because it reduces the gap between real and virtual environments, on the other hand, representing an entire façade of a building in one frame is not realistic. Thus, it requires additional complexity, since we have to define also a support for a vertical scroll of the content, to show the not visible parts of the frame.

More in details, the development of NUIs to explore the content of wall-sized displays with the gaze, requires two subtasks:

1. Defining techniques to estimate the PoG of the user while he/she is looking at the display, and
2. Defining a navigation logic associated to the PoG.

In the following, we provide technical details on how we faced these two tasks.

2.1 Point of gaze estimation

Head poses are usually computed by considering 3 degrees of freedom (DoF) [17], i.e. the rotations along the 3 axis of simmetry in the space, *x, y,* and *z*, shown in Figure 2.

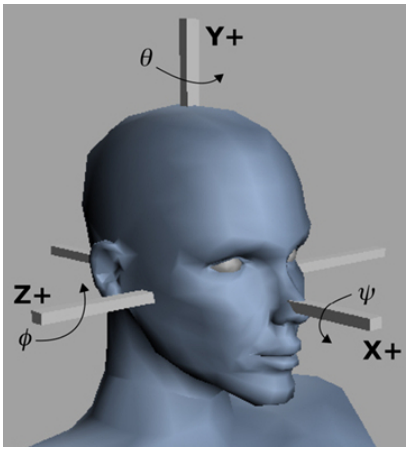

Figure 2: Head movements.

Once the head pose in the space is known, the pupil center position can optionally refine the PoG estimation.For example, in the medical diagnosis scenario, to estimate the PoG, patients are usually not allowed to move their head [7] or they have to wear head-mounted cameras pointed towards their eyes [12]. In these cases, to estimate the PoG means to compute the pupil center position with respect to the ellipse formed by the eyelids, while the head position, when considered, is detected through IR sensors mounted on the head of subjects. These systems grant an error threshold lower than 5 pixels [12], achievable thanks to strict constraints on the set-up, such as the fixed distance between eye and camera but, on the other hand, they have a very high level of invasiveness for the users. In other scenarios, the PoG is estimated by means of remote trackers, such as ones presented in [6], which determine the gaze direction by the head orientation. These systems do not limit users' movements and do not require them to wear any device.

In the cultural heritage context, the gaze detection is mainly used for two tasks. The first one is related to the artistic fruition: according with "The More You Look The More You Get" paradigm [16], users focusing their gaze on a specific work of art or part of it, can be interested to receive some additional content about that specific item. This usage of the gaze direction can be extremely useful in terms of improving the accessibility to the cultural heritage information and enhancing the visit experience quality. The second task is related to understanding how people take decisions, visiting a museum: which areas they are focused and how long; outputs from gaze detectors are then gathered and analyzed [18].

Starting from an approach we already developed for small displays (between 50 x 30cm and 180 x 75cm) [4], we propose an extension for wall-sized ones, based on a combined exploitation of the head pose and pupil size to explore digital environments. The general settings of the display is presented in figure 3. In particu-

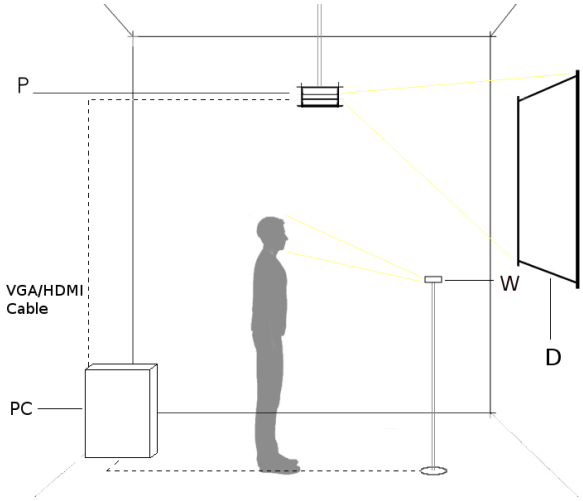

Figure 3: Gaze detection: experimental settings.

lar, the exhibition set up includes a *PC* (the machine on which the software runs), a webcam *W* which acquires the input stream, and a projector *P* which beams the cultural content on the wall-sized display *D*. We assume the user to stand almost centrally with respect to *D* and with a frontal position of the head with respect of the body.

In the previous work with small displays [4], we used an eye-tracking technology to estimate the gaze, since we experienced that, for limited sizes, users just move the eyes in order to visually explore the surface of the artwork. On the other hand, in the case of wall sized displays, users have to move also their head, performing thus limited ocular movements.

Therefore, an head pose estimator is needed. To this, according to related work [8], we developed a solution aimed at tracking the nose tip of the user in 3 Degrees of Freedom (DoF). Indeed, the nose tip is easy to detect and, since it can be considered as good approximation of the head centroid, given the required precision from our domain, it is a useful indicator of the head position in the three-dimensional space.

2.1.1 Nose Tip detection

The first step in the processing pipeline is to detect, within the video stream from the webcam, the face of the user. According to the literature, this task can be executed with different strategies, which can be grouped in two main sets: the image-based, such as skin detection [10], and the feature-based. In our approach, the detection of the face is based on a solution from the second group, namely the Haar feature-based Viola-Jones algorithm [24]. In a first implementation, we scanned the entire image to locate the face; subsequently this search was improved, providing as input the range of sizes for a valid face, depending on the distance between user and camera.

Within the area of the face, also the nose tip search is performed by means of the Viola-Jones algorithm, in terms of its OpenCV implementation, which returns the nasal area centered on its tip. Initially, we searched for the nose scanning the entire face; then, we considered that the search could be improved by taking advantage of the facial geometric constraints [13], to increase both precision and computational efficiency. In particular, the nose can be easily found starting from the facial axis on y axis and from the middle point of the face, for both x and z axis. We performed the search on images of size 1280 x 960 pixels, processed on an Intel Core i7 with 2.2 GHz; initially, the detection time was about 100 ms. The optimizations on face and nose search allowed us to locate the face and the nose on average in 35 ms, reducing the computation time of about 65%.

2.1.2 Nose Tip tracking

The previously described features are searched either the first time a user is detected or when the tracking is lost. In all the other frames, the nose tip is simply tracked.

Several strategies have been proposed to track the motion, that can be categorized into three groups: feature-based, model-based and optical flow-based. Generally speaking, the feature-based strategies involve the extraction of templates from a reference image and the identification of their counterparts in the further images of the sequence. Some feature-based algorithms need to be trained, for example those based on Hidden Markov Models (HMM) [21] or Artificial Neural Networks (ANN) [14], while others are non-supervisioned, like for instance the Mean Shift Tracking algorithms [28]. Although the model-based strategies could be considered a specific branch of the feature-based ones, they require some a-priori knowledge about the investigated models [27]. The optical flow is the vector field which describes how the image changes during the time; it can be computed with different strategies as, for example the gradient.

In our approach, we adopted a non-supervisioned feature-based algorithm. Thus, we firstly store the image region containing the feature (i.e. the nose tip), to be used as template. Then, we apply the OpenCV method to find a match between the current frame and the template. The method scans the current frame, comparing the template image pixels against the source frame and stores each comparison result in the resulting matrix. The source frame is not scanned in its entirety, but only a Region of Interest (ROI) has been taken into account; the ROI corresponds to the area around the template coordinates in the source image. The resulting matrix is then analysed to find the best similarity value, depending on the matching criterion given as input. We used the *Normalized Sum of Squared Differences* (NSSD) as matching criterion, whose formula is reported in equation 1.

$$R(x,y) = \frac{\sum_{x',y'} (T(x',y') - I(x+x',y+y'))^2}{\sqrt{\sum_{x',y'} T(x',y')^2 \sum_{x',y'} I(x+\dot{x}',y+y')^2}} \quad (1)$$

In equation 1, T is the template image and I is the input frame in which we expect to find a match. The coordinates *(x,y)* represent the generic location in the input image, whose content is being compared to the corresponding pixel of the template, located at *(x',y')*. *R* is the resulting matrix and each location of *(x,y)* in *R* contains the corresponding matching result. The minimum values in *R* represent the minimum differences between input image and template, indicating the the most likely position of the feature in the image. Thus, while a perfect match will have a value of zero, a mismatch will have a larger sum of squared difference. When the mismatch value exceeds the confidence level [19], the tracking is lost.

2.2 Projecting the Nose Tip for Navigation

Our second task is associating an action to the gaze. To this aim, we have to understand where the user is looking at, on the wall-sized display. Since we can approximatively interpret the nose tip as centroid of the head, in order to provide a coherent PoG estimation, we have to solve the proportion to transpose the nose tip coordinates into the display reference system. To this aim, we ge-

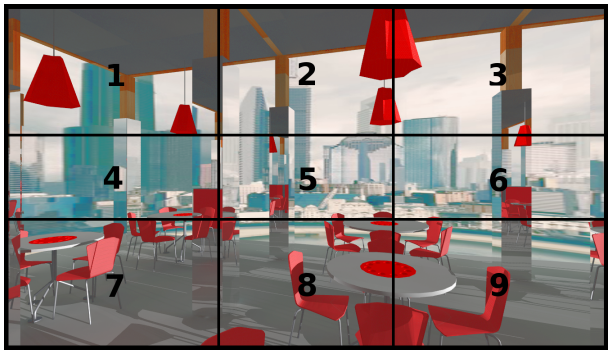

Figure 4: Matrix Model of the Wall-Sized Display.

ometrically project its coordinates on the observed wall-sized display reference system. These new coordinates are calculated and then tracked with respect to the shown frame. The area of the wall-sized display is considered as a 3x3 matrix, as shown in figure 4. What we do in the current implementation is to indicate in which cell of the matrix the gaze is falling.

When the user stands in front of the display with the head centered in frontal position, the geometric projection of his/her nose tip falls into the cell #5 of the matrix (2nd row, 2nd column). We defined the size of the central row to obtain a kind of "comfort zone", where minor movements of the head are not triggering any movement of the rendered image. In details, head rotations up to 15 degrees on the x axis and up to 8 degrees on both the y and the z axes do not affect the gaze position. With wider rotations, the projection of the nose falls in another cell, and the digital image will be shifted accordingly.

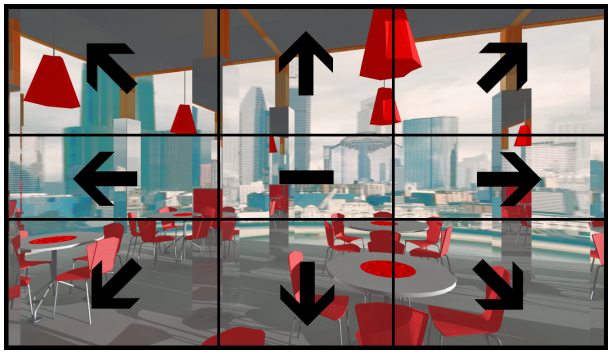

Figure 5: Input actions associated with the gaze directions.

According to the <*event, condition, action*> paradigm [22], the *event* is the identification of a fixation point; the *condition* is marked by the index of the cell in the 3x3 matrix and the corresponding *action* is defined in figure 5. In particular, as explained in figure 5, when the PoG falls in the cells #4 or #6, we associate the action of navigating the content on the left side or on the right side, respectively. When the user observes the sections #2 or #8, the content will be navigated upwards or downwards; the section #5 will be interpreted as the area in which no action will be executed. When the PoG falls in the remaining cells, the content will be navigated in the respective diagonal directions.

In the current implementation, since we are just associating a cell of the matrix to the PoG, the speed of the scroll is fixed and independent from the PoG of the user within a lateral cell of the matrix. We are currently implementing a new version of the navigation paradigm, where this 3x3 matrix will be replaced by a continuous function, where the speed of the scroll will be proportional to the distance of the POG from the center of the display.

3. THE EMOTIONAL CONTRIBUTE

One of the problems with NUIs based on the PoG estimation is that it is difficult to understand the reaction of the user in terms of interest towards the shown content [26].

To address this issue, we developed a further video processing module, intended as a complement to the system presented in the previous section, and able to detect implicit user feedbacks. The output of this module can be used for a twofold objective: it could trigger in real-time reactions from the system, and/or it can provide a powerful post-visit tool to the curator of the exhibition, with a log of the reactions of the visitors to the shown digital content. In this way, the curator could get a better insight on the content which is sparkling the highest interest in the visitors. In the following we provide some technical details on how we faced this issue.

3.1 The Mydriasis

A wide range of medical studies proved that the brain reacts to the emotional arousal with involuntary actions performed by the sympathetic nervous system (e.g.: [9] [11]). These changes manifest themselves in a number of ways, such as increased heart-beat, higher body temperature, muscular tension and pupil dilation (or *mydriasis*). Thus, it could be interesting to monitor one or more of these involuntary activities to discover the emotional reactions of the visitors while they are enjoying the cultural contents, in order to understand which details arouse pleasure.

In the age of wearable devices, there are many sensors with health-oriented capabilities, like for instance armbands or smartwatches, that could monitor some of these involuntary actions of our body. For instance, information about the heart-beat or the body temperature can be obtained by means of sensors which retrieve electric signals, once they are applied on the body. If by one side these techniques grant an effective level of reliability, on the other side they could influence the expected results of the experiments, as users tend to change their reactions when they feel under examination [11]. Moreover, they would require the visitors to wear some special device (having also high costs for the exhibition), which could be a non-viable solution in many contexts. For these reasons, we again looked for a remote solution, able to get an insight on the emotional arousal of the visitor without requiring them to wear any device.

Given the set-up described in Section 2.1, we tried to exploit additional information we can get from the video stream collected by the webcam. In particular, we tried to remotely monitor the pupils behaviour during the interaction with the wall-sized display. Let us note that, as both pupils react to stimuli in the same way, we studied the behaviour of one pupil only.

Pupils are larger in children and smaller in adults and the normal size varies from 2 to 4 mm in diameter in bright light, and from 4 to 8 mm in the dark [25]. Moreover, pupils react to stimuli in 0.2 s, with the response peaking in 0.5 to 1.0 s [15]. Hess presented 5 visual stimuli to male and female subjects and he observed that the increase in pupil size varied between 5% and 25% [9].

3.2 Pupil detection

Before detecting the pupil, we have to locate and track the eye on the video stream coming from the webcam. The detection is performed by means of the Haar feature-based Viola-Jones algorithm [24], already cited in section 2.1.1, while the tracking of the pupil is done with the template matching technique, as described in

section 2.1.2.

The detected ocular region contains eyelids, eyelashes, shadows and light reflexes. These represent noise for pupil detection, as they could interfere with the correctness of the results. Thus, the eye image has to be pre-processed, before searching for the pupil size. We developed a solution including the following steps, in order to perform the pre-processing:

1. The gray scaled image (Figure 6a) is blurred by means of a median filter, in order to highlight well defined contours;
2. The Sobel partial derivative on the x axis reveals the significant changes in color, allowing to isolate the eyelids;
3. A threshold filter identifies the *sclera*.

As result, these steps produce a mask, which allows us to isolate the eye ball from the source image. Pupil detection is now performed on the source image as follows:

1. We drop down to zero (black) all the pixels having cumulative distribution function value greater than a certain threshold [1] (Figure 6b);
2. We morphologically transform the resulting binary image by means of a dilation process, to remove the light reflexes on the pupil;
3. A contours detection operation identifies some neighbourhoods (Figure 6c).
4. The pupillary area is found by selecting the region having maximum area (Figure 6d);
5. The center of the ellipse (Figure 6e) best fitting the pupillary area, approximates the pupil center (Figure 6f).

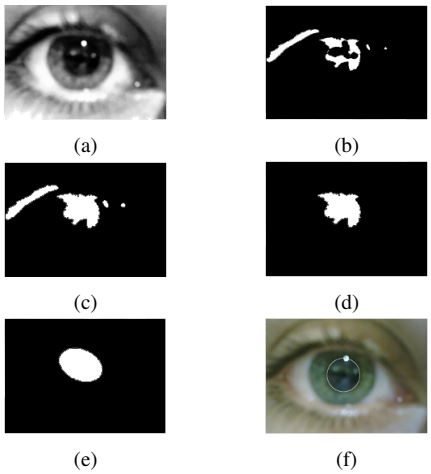

Figure 6: Pupil processing steps.

Once we detected the pupil, to calculate the mydriasis we store the first computed radius and, frame by frame, we make a comparison between the first radius and the ones calculated during the following iterations: according to Hess, when the comparison exceeds the 5%, a mydriasis is signaled.

To log all these implicit feedbacks, during the interaction a parallel thread keeps track of the observed sections and the related emotional reactions. In particular, at fixed steps of 200 ms, the thread saves the current timestamp, the index of the observed section and an integer value representing the pupil status. If the pupil has normal size, the pupil status is 0, otherwise it is 1. If the system does not detect a face for a given time (10 seconds, in the specific) the interaction session is considered terminated and the collected information is stored in a a XML document. The structure of the XML document is shown in the Listing 1.

Listing 1: A snippet of the logging file

```
<?xml version="1.0" encoding="UTF-8"?>
<reportCollection>
<report id = "0">
  <track idTs="1402674690300" section="
      -1" mydriasis ="0" />
  <track idTs="1402674690500" section="1
      " mydriasis ="0" />
  <track idTs="1402674690700" section="1
      " mydriasis ="0" />
  <track idTs="1402674690900" section="1
      " mydriasis ="0" />
  <track idTs="1402674691100" section="1
      " mydriasis ="0" />
</report>
<report id = "1">
  <track idTs="1402675341320" section="1
      " mydriasis ="0" />
  <track idTs="1402675341520" section="0
      " mydriasis ="0" />
  <track idTs="1402675341720" section="0
      " mydriasis ="0" />
  <track idTs="1402675341920" section="0
      " mydriasis ="0" />
</report>
</reportCollection>
```

The XML document is created and initialized with an empty <reportCollection>, when the application starts; then, when each interaction session ends, a new <report> subtree is created. The timestamps values univocally identify the respective <track> elements. Given, this simple structure, it is easy to perform subsequent analyses of the interaction session of the visitors.

4. THE CASE STUDY

The system we developed was shown at the Tianjin International Design Week 2015, for the personal exposition dedicated to the Italian designer Matteo Ragni. In particular, the software was used to let the visitors to navigate with the gaze the 360°virtual reconstruction of Matteo Ragni's *Camparitivo in Triennale*, on a wall sized display. In order to implement the case study, we started from the design model of *Camparitivo in Triennale*, in Rhino3D format [3], including the textures obtained from photos, and we placed a virtual camera into the center of the model, to have the point of view of a visitor inside the *Camparitivo*. With this settings, we rendered a complete rotation of the camera around a fixed vertical axis corresponding to the imaginary *neck* of the visitor, in order to obtain photorealistic, raytraced reflections on the mirrors. With this setup, we obtained 360 images with a step size of 1 degree. An illustrative frame is shown in Figure 7. We considered each frame as divided according to the matrix in figure 4. Once the system indicated the observed section of the matrix, the respective action of figure 5 was executed and the related frame was shown.

4.1 The experiments

[3]www.rhino3d.com

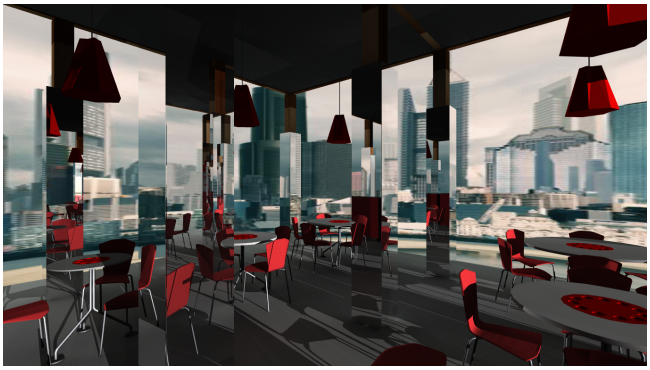

Figure 7: A frame of the rendered model.

Figure 8: The experimental setting.

Basically, motor tasks such as "look at there" are performed in video games by hands controlled operations, because they are usually executed by classical input devices such as: joystick, joypad, keyboard or mouse. Our work represents an attempt to improve the naturalness of this kind of interaction, by associating the task with its implicitly corresponding interface. We left the users free of interacting with the application, without giving them any kind of instruction or support. The only source of information for them was represented by the panel shown in figure 8, explaining that the input was given by the head movements and not by the eyes.

During the exposition, more than 150 visitors experienced our stand, standing at 1 meter from a webcam mounted at 160cm of height, as shown in Figure 8. Among all the visitors, 51 speaking English accepted to answer to a quick oral interview, as we could not submit written questionnaires during the public event for logistic reasons.

After we asked users if it was the first time they experienced

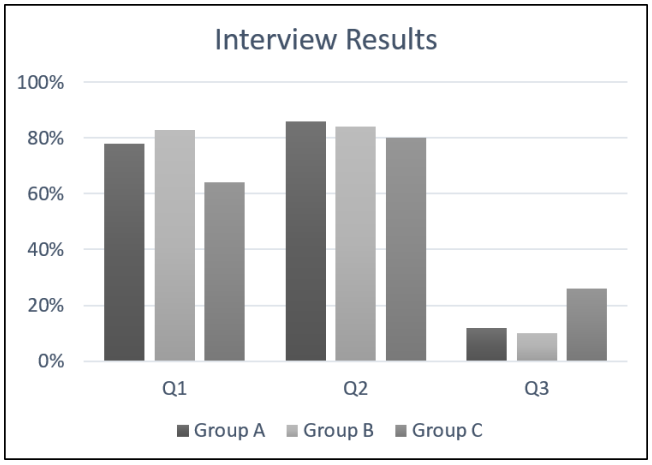

Figure 9: Cumulative Results of the Interviews

a gaze-based application, we submitted the following questions to them:

1. Do you think this kind of application is useful to improve the museum fruition?
2. Did you find the application easy to understand?
3. Did you find any difficulties during the interaction?
4. How old are you?

Participating subjects were grouped in three subsets, according to their age, where all the subsets have the same number of subjects. The group A has users whose age is between 18 and 35 years; group B corresponds to people from 36 to 65 years old; group C is composed by users older than 65 years. We did not make distinction between male and female subjects. For all of them, it was the first time they tried a gaze-based IT solution.

4.2 Results

The results of this very preliminary evaluation of the proposal are reported in Figure 9, where the histograms represent the percentage of positive answers given by the subjects over the total of answers. Please note that for Q1 and Q2, the higher the results, the better is the feedback, while for Q3, the lower the better.

Interpreting the comments of the users, as for Q1, we see that the vast majority of the subjects believe the proposed interface was useful to improve the cultural experience. People older than 65 are less enthusiastic, but this is somehow an expected result. As for Q2, an even higher percentage of subjects found the application easy to understand. For Q2 there is less difference among the three groups. Finally, as for Q3, we found that some of the subjects encountered difficulties in interacting with the software, with a significant difference for the Group C with respect to the other two groups. In general problems arose when visitors performed rapid or wide head movements. In both cases, this led to a failure of the nose tip tracker. In particular, when the users performed wide rotations, the template matching results exceeded the confidence level, causing the lost of the tracking. Similarly, rapid head movements caused a sudden reduction of the similarity between frame and template, causing the tracker to fail.

An objective survey about the user experience has been conducted by analyzing the collected log data. In particular, we used the stored timestamps and the indexes of the observed Regions Of

Interest, to indicate the duration of each interaction and on which regions users concentrated their gaze. Data showed that 45% of users performed a complete interaction, observing all 9 ROIs. According to the matrix in Figure 4, the most observed ROI has been the #4, observed by 88%of users. The average duration of the interaction has been 95 seconds per user.

All in all, we can see from this very preliminary investigation that visitors largely enjoyed the experience with the gaze-based interaction.

As for the mydriatic reactions of users, this is more problematic. We analyzed the logs of the exhibition, and we found that the mydriatic reactions occurred in:

- 65% of cases for group A;
- 40% of cases for group B;
- 20% of cases for group C.

There are two consideration to drawn from these numbers. The first is that in general the technological solution is not mature enough for a wide public. This is particularly true for Asiatic people, as the totally of the subjects had black eyes, which makes the identification of the pupil more problematic. Some internal investigations we did with Caucasian subjects led to better results. The other conclusion is that there is a well-known difference in the mydriatic reactions with respect to the age of the subjects, where the older they are, the smaller are the differences in the size of the pupil between the relaxed and aroused states. So, it is clear that the emotional module requires further research efforts.

5. CONCLUSIONS

Wall-sized displays represent a viable solution to present artworks difficult or impossible to move. In this paper, we proposed a Natural User Interface to explore 360°digital artworks shown on wall-sized displays, allows visitors to look around and explore virtual worlds using only their gaze, stepping away from the boundaries and limitations of the keyboard and mouse. We chose to accomplish the task by means of a remote head pose detector. As it does not require calibration, it represents an immediate to use solution for supporting digital environment navigation. Moreover we developed a solution to monitor the mydriatic reactions of the subjects while they were using the system, to get an implicit feedback on the interested of the represented digital content. A preliminary investigation we performed at the Tianjin International Design Week 2015 with 51 subjects gave us the feedback that the gaze-based navigation can be well-accepted by the visitors, as it is felt as a way to improve the fruition of Cultural Heritage. Nevertheless, the monitoring of mydriatic reactions should still be improved, especially for people with black eyes.

Anyhow, from the results we collected, there are still many potential research direction for this topic. First of all, we are currently developing a new version of the system where the display is no more divided into a matrix, but instead there will be a smooth feedback from the system, whose rapidity of response will be more proportional correlated to the amount of movement done by the head of the user. The second main research field is to extend this approach towards freely explorable 3D environments, thus to support also the forward and backward navigation. The idea of enrich gaze with forward and backward navigation has been approached in different works. One solution is the *fly-where-I-look* [2] in which authors associate the interest of users to *fly* towards an area, with the action to look at it. This approach finds basis in cognitive activities: in particular, some studies prove that more fixations on a particular area indicate that it is more noticeable, or more important to the viewer than other areas [20]; Duchowski [5] estimates the mean fixation duration of 1079 ms. This approach represents a natural and simple solution to the task "look forward", but the activation time forces the user to wait for the operation starts, without doing anything and it may feel like a waste of time. Finally, also voice commands could be a natural input to perform this task; thus, our current research direction is oriented to provide a better support for multimodal interaction.

6. ACKNOWLEDGEMENT

This work has been partly supported by the European Community and by the Italian Ministry of University and Research (MIUR) under the PON Or.C.He.S.T.R.A. (ORganization of Cultural HEritage and Smart Tourism and Real-time Accessibility) project.

7. REFERENCES

[1] M. Asadifard and J. Shanbezadeh. Automatic adaptive center of pupil detection using face detection and cdf analysis. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, page 3, 2010.

[2] R. Bates, H. Istance, M. Donegan, and L. Oosthuizen. Fly where you look: enhancing gaze based interaction in 3d environments. *Proc. COGAIN-05*, pages 30–32, 2005.

[3] D. M. Calandra, A. Caso, F. Cutugno, A. Origlia, and S. Rossi. Cowme: a general framework to evaluate cognitive workload during multimodal interaction. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 111–118. ACM, 2013.

[4] D. M. Calandra, D. Di Mauro, D. D'Auria, and F. Cutugno. Eyecu: an emotional eye tracker for cultural heritage support. In *Empowering Organizations*, pages 161–172. Springer, 2016.

[5] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[6] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 617–624. IEEE, 2011.

[7] E. S. Gómez and A. S. S. Sánchez. Biomedical instrumentation to analyze pupillary responses in white-chromatic stimulation and its influence on diagnosis and surgical evaluation. 2012.

[8] D. Gorodnichy. On importance of nose for face tracking. 2002.

[9] E. H. Hess and J. M. Polt. Pupil size as related to interest value of visual stimuli. *Science*, 132:349–350, Aug. 1960.

[10] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[11] D. Kahneman and J. Beatty. Pupil diameter and load on memory. *Science*, 154(3756):1583–1585, 1966.

[12] M. Kassner, W. Patera, and A. Bulling. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. April 2014.

[13] T. T. Le, L. G. Farkas, R. C. Ngim, L. S. Levin, and C. R. Forrest. Proportionality in asian and north american caucasian faces using neoclassical facial canons as criteria. *Aesthetic plastic surgery*, 26(1):64–69, 2002.

[14] H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. *Image Processing, IEEE Transactions on*, 9(1):147–156, 2000.

[15] O. Lowenstein and I. E. Loewenfeld. The pupil. *The eye*, 3:231–267, 1962.

[16] S. Milekic. The more you look the more you get: Intention-based interface using gaze-tracking. 2003.

[17] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.

[18] R. NETEK. Implementation of ria concept and eye tracking system for cultural heritage. *Opgeroepen op september*, 9:2012, 2011.

[19] K. Nickels and S. Hutchinson. Estimating uncertainty in ssd-based feature tracking. *Image and Vision Computing*, 20(1):47 – 58, 2002.

[20] A. Poole, L. J. Ball, and P. Phillips. In search of salience: A response-time and eye-movement analysis of bookmark recognition. In *People and Computers XVIII—Design for Life*, pages 363–378. Springer, 2005.

[21] L. R. Rabiner and B.-H. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[22] B. Shneiderman. *Designing the user interface*. Pearson Education India, 2003.

[23] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2012.

[24] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001.

[25] C. VL and K. JA. Clinical methods: The history, physical, and laboratory examinations. *JAMA*, 264(21):2808–2809, 1990.

[26] D. Wigdor and D. Wixon. *Brave NUI world: designing natural user interfaces for touch and gesture*. Elsevier, 2011.

[27] P. Wunsch and G. Hirzinger. Real-time visual tracking of 3d objects with dynamic handling of occlusion. In *Robotics and Automation, 1997. Proceedings., 1997 IEEE International Conference on*, volume 4, pages 2868–2873. IEEE, 1997.

[28] C. Yang, R. Duraiswami, and L. Davis. Efficient mean-shift tracking via a new similarity measure. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 176–183. IEEE, 2005.