Cultural heritage presentations with a humanoid robot using implicit feedback

Antonio Origlia
Dept. of Electrical Engineering and Information Technology, University of Naples "Federico II"
Naples, Italy
antonio.origlia@unina.it

Antonio Rossi
Dept. of Electrical Engineering and Information Technology, University of Naples "Federico II"
Naples, Italy
rossi.antonio.84@gmail.com

Maria Laura Chiacchio
Dept. of Electrical Engineering and Information Technology, University of Naples "Federico II"
Naples, Italy
marialaurachiacchio@gmail.com

Francesco Cutugno
Dept. of Electrical Engineering and Information Technology, University of Naples "Federico II'
Naples, Italy
cutugno@unina.it

ABSTRACT

In recent years, there has been an increasing interest towards cultural heritage in the field of ICT applications. To design efficient communication strategies the knowledge possessed by art historians, with expertise in mediating access to cultural heritage, has become a valuable resource. In this work, we present a human-robot interaction setup where people actively choose how much information they would like to access concerning the available topics. To provide engaging presentations, a humanoid robot exhibiting a general behaviour based on a human presenter was used and a mathematical model to keep track of content navigation was designed. Monitoring the evolution of the interactive session allows to estimate users' general interest towards the available contents. Our results show that people were very satisfied by the interaction experience and that automatically detected interests were consistent with the users'. Both subjective and objective metrics were used to validate the approach.

CCS Concepts

•Human-centered computing → Collaborative interaction;

Keywords

Cultural heritage presentation, human-robot interaction, implicit feedback

1. INTRODUCTION

Communication in museums is considered an important issue even if museum specialists have sometimes been reproached for not doing enough in this field. Many advancements have been obtained in the last years concerning the attempt to understand museum visitors needs and develop new interactive ways to address these. Among many others, investigations about visitors psychological approach [**?**] helped museologists to develop possible methods not only to exhibit artifacts but also to give them a sense, providing further explanations. So museums experts may be ready to take an active role in the development of technologies to support visitors' experiences [**?**]. Artificial agents can support visitors to obtain information about cultural heritage in a pleasant way. This is because they use interaction paradigms that do not require users to deviate too much from everyday communication. Of course, an artificial agent must rely on written content as Natural Language Generation techniques are still experimental. To obtain a resource enabling an artifact to *talk* to a user, it is important to start from texts as close as possible to spoken language, as previous investigation in psychology suggests, too [**?**]. In order to obtain such a resource, we compiled our reference database starting from speech transcriptions. We collected speech material of a human expert presenting works of art and converted it into a form an artificial agent can use in a presentation task. There are a number of examples of robots being used in cultural heritage dissemination. Among existing systems for interactive edutainment with robots, ROBOTINHO [**?**] provides multimodal interaction through spoken presentations, facial expressions and gestures. Its hands convey greetings, deictic and other spatial information, while gaze establishes joint attention. Also, the robot presented in [**?**] attracts visitors attention through verbal and nonverbal action. In these works, the authors concentrated on the development of social strategies to obtain a believable artificial presenter. In interactive approaches, user feedback has been used to obtain user models for recommendation systems and personalisation (e.g. [**?**]). Most of these systems, however, use explicit

feedback, typically rating, to recommend items. However, research in the Information Retrieval field highlighted that explicit feedback poses a significant problem: the obtrusiveness of the approach [**?**]. This problem becomes critical in edutainment setups.

In this paper, we describe how the robot's behavioural strategies and discourse structure were designed on the basis of orally delivered cultural heritage presentations. From the transcription of these recordings, a series of information nodes have been identified and organized in a tree structure representing general concepts, topic sharing and deepening levels. This structure is then automatically populated with information describing the feedback strength the system should consider, depending on how data are explored. An interactive task using a humanoid robot and a tablet interface is used for validation.

2. ROBOT BEHAVIOUR DESIGN

To correctly design the robot's behaviour during presentations, we collected a corpus of audio-visual reference material to study how a human expert delivers the contents related to the considered works of art. In this case, one of the authors (Maria Laura Chiacchio) performed the presentations. The total recorded material consists of two hours and a half. In this work, we concentrate on reproducing speech and gestures in a humanoid robot, but the full dataset contains material for future analyses.

Considering that every attempt to classify art may be controversial, it has been anyway necessary for the present research to look for wide categories that could be representative of the most famous styles of European Art. In each category, among many famous artists and masterpieces, the choice of the paintings has been done considering their importance towards the category itself. Moreover, it has been important also to avoid very well-known masterpieces, because their fame could influence the choice of the users to go further with the exploration. So, for example, instead of considering the *Gioconda* by Leonardo da Vinci, another of his paintings has been presented, which is a famous work as well but not a universally recognized icon.

The high-quality recorded speech was automatically transcribed and manually corrected to remove disfluencies and filled pauses. Punctuation and Synthetic Speech Markup Language (SSML) tags were also added to support the generation of synthetic speech, obtained using the MIVOQ[1] engine. To make the robot move consistently, a set of gestural strategies to accompany the presentations were observed and used to control the robot's behaviour:

- When the presentation refers to general aspects of the painting, the presenter looks at it to attract the listener's attention to the work of art;
- When the presentation refers to a specific detail of the painting, the presenter points at the area where the detail is found. This is because it is not always straightforward, for the listener, to identify the specific point where the presentation subject can be found. Using gestures is more immediate and precise than using spoken instructions about where to look;
- When the presentation refers to a specific area of the painting, the strategy is similar to the one used in the preceding point. However, a different gesture, encircling the area of interest, has been identified as different from simple pointing;
- When the subject of the presentation is not directly related to the painting (e.g. it refers to the author's life) the presenter looks at the audience. This signals the absence of direct correlates of the presentation in the painting.

[1]www.mivoq.it

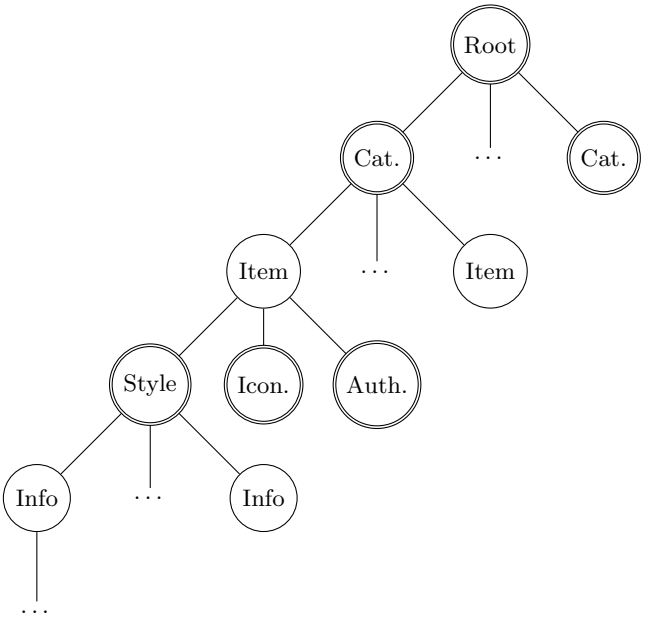

Figure 1: General structure of the XML tree representing contents organization.

3. COLLECTING IMPLICIT FEEDBACK

The general structure of the data is organized into an XML tree where two kinds of nodes are available: *data* nodes and *abstract* nodes. The tree is composed by the following elements:

- An abstract root node
- A number of abstract *Category* nodes grouping paintings belonging to the same art movement
- a number of data *Item* nodes containing a short presentation of the painting considered in the relative subtree
- for each *Item* node, three abstract children nodes are considered to group data concerning *Style, Iconography* and *Author*
- for each of these abstract nodes, there is a subtree of data *Info* nodes containing SSML. In this subtree, a child-parent relationship indicates that the child node contains a deepening of the content provided in the parent node. Sibling nodes provide different insight about the topic covered by their parent.

A graphical summary of the data structure is presented in Figure 1. In order to evaluate the interest of a user towards a specific category, it is necessary to assign a measure

of importance to each node in the subtree rooted for every provided category. The concept is that interest is maximal when all the representative items of the category are fully explored by the user. Since we assume that requesting a topic deepening brings more insight into the user's interests, it is important to keep track of the structure of each considered subtree. Specifically, we consider the *resolution* $R(T)$, of a generic T tree, which corresponds to the number of *internal branches*, branches that do not connect a node to a leaf [**?**]. We also define the number of nodes included in a tree T as $N(T)$ and the *informative potential* $I_p(T)$ as the potential interest the user may implicitly express by exploring the T tree. Abstract nodes entirely distribute the amount of informative potential they receive among their subtrees. Data nodes retain a certain part of the *informative potential* before passing the remaining amount to the subtrees rooted into their children. We will refer to the amount of the *informative potential* retained by the n_{th} data node as its *informative content* $I_c(n)$. Also, we will refer to the amount of *informative potential* distributed to children nodes as its *informative residual* $I_r(n)$. $I_c(n)$ is computed by retaining the fraction of *informative potential* that would be assigned to it, should it be equally distributed among the nodes in its descendant nodes, as shown in Equation 1

$$I_c(n) = \frac{I_p(T_n)}{N(T_n)} \tag{1}$$

where T_n is the tree rooted in the n_{th} node. Consequently, $I_r(n)$ is computed as shown in Equation 2

$$I_r(n) = I_p(T_n) - I_c(n) \tag{2}$$

The informative potential of the tree rooted in a *Category* node is assigned a value of 1: this potential has to be distributed among the subtrees containing *Item* nodes. As different *Items* have different amounts of information and different structural organization, it is necessary to distribute the informative potential among the subtrees in such a way that the interest value of parenthood is more powerful than the one of siblinghood. The strategy to distribute the *informative residual* to the subtrees depends on two factors: the number of nodes included in a subtree and the resolution of the subtree itself. Half the score depends on the former while the other half on the latter. One half of the *informative residual* is weighted by the ratio between the resolution of the considered subtree and the sum of the resolutions of the considered subtree and the resolutions of the trees rooted in the siblings of its root. The other half is weighted by the ratio between the number of nodes in the considered subtree and the total number of nodes in the tree itself and in its siblings. $I_p(T)$ is therefore computed as in Equation 3

$$I_p(T) = \frac{I_r(p)}{2} \cdot \left(W_r(T) + \frac{N(T)}{\sum_{i=1}^{N_s} N(T_i)} \right) \tag{3}$$

where $W_r(T)$ represents the portion of informative residual that is assigned to the T subtree on the basis of its resolution. Of course, if the sibling nodes are all roots of subtrees with resolution equal to 0, the entire informative residual is assigned depending solely on the number of nodes, as shown in Equation 4

$$W_r(T) = \begin{cases} \frac{R(T)}{\sum_{i=1}^{N_s} R(T_i)}, & \text{if } \sum_{i=1}^{N_s} R(T_i) \geq 1 \\ \\ \frac{N(T)}{\sum_{i=1}^{N_s} N(T_i)}, & \text{otherwise} \end{cases} \tag{4}$$

where N_s is the number of siblings of the root of T plus 1 (the root itself) and p is the parent node of the root of the T tree.

This approach is motivated by the way paintings are presented by the virtual agent: first of all, it provides access to the content of the root node of the considered *Item*, which contains generic data about the presented work of art. When the introduction is finished, the user is allowed to choose if continuing to navigate the tree or move to another *Item*. In general, after providing the contents of each accessed node, the user is given three choices:

- if at least one child node is available, the user can request a *Deepen* move. This will make the virtual agent access the contents of its first child;
- if at least one unvisited sibling is available, the user can request a *Continue* move, which will make the virtual agent access the contents of the first sibling of the current node;
- the user can always choose to *Exit* to the next *Item*.

Given the value of the parenthood relationship a *Deepen* move can be considered to be totally informed as the user is aware that the chosen action will bring more information about the topic it just listened to. This can be considered a strong indication of interest for the topic. A *Continue* move is less informed because, while the user is aware that the topic will insist on the current painting, she cannot predict the exact topic. The virtual agent automatically moves to the next *Item* if the current tree is exhausted. We consider the total interest score to be the sum of the *informative content* of the visited nodes.

4. RESULTS

In order to evaluate the overall quality of the presented method and interest detection in art categories, we used a human-robot interactive setup made of an Aldebaran Nao unit paired with a tablet. The robot uses a synthetic voice interpreting SSML data to mimic the expressive style of the recorded expert. It also controls the tablet to show the paintings and to zoom in areas of interest tied to specific nodes of the XML presentation tree. The tablet interface allows users to issue *Continue*, *Deepen* or *Next* moves. A group of 30 users was recruited and provided with written instructions describing the interface. The subjects were people from the university with competence in computer science. The categories used for the experiments are *Renaissance, Impressionism* and *Avant-garde*. An average user session lasted between 20 and 30 minutes. Users were asked to explore the presentations and look for a painting they would be interested in discussing. After the interaction session was over, users chose and commented a single painting and evaluated, on a scale of 1 to 5, the quality of the presentation offered by Nao. Asking users to pick a single painting instead of a specific category is meant to obfuscate the goal of the experiment and to simplify the task for non-expert

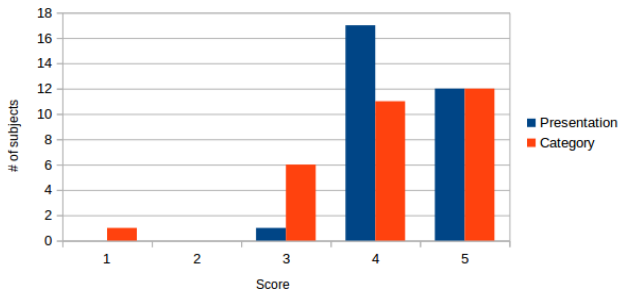

Figure 2: Distribution scores for presentation quality and category selection.

users. It makes more sense, for a user, to select an interesting item in a set they just explored as they may be unaware of the art movement the item belongs to. The performance measure is then given by the agreement rate between the category selected by Nao, which is kept hidden to the users at this stage, and the category to which the chosen painting belongs to. To collect an explicit judgement on this, at the end of the experiment users were informed about which category scored the highest value and they were provided with a brief, written description of the category that was considered to be of their potential interest. Users were then asked to evaluate the automatic choice on a scale of 1 to 5.

First of all, we consider explicit judgement given by the users. The obtained distribution of scores for presentation quality and category selection is shown in Figure 2. Results show that the subjects assigned a significantly high score to the overall experience offered by Nao, validating the transfer of the expert's basic behaviour in the robot. This is important as we can safely discard the possibility that indications of non-interest were caused by the robot. Explicit judgement for category selection is also high, on average. A single, very low, outlying score was observed from a person who declared that he "*did not like art in general*".

To evaluate the agreement rate between the categories of the paintings chosen by the subjects and the categories selected by Nao, we consider Weighted Cohen's Kappa. In general, this measure evaluates the agreement rate between two annotators. In our case, we also need to differently weight errors given the relationship between the considered categories. We do this by specifying the relative distances between the categories and then applying squared weighting as a function of the temporal ordering from *Renaissance* to *Avant-garde*. More in detail, *Impressionism* and *Avant-garde* are close as the former poses the basis for the latter. Impressionism is closer to Renaissance than *Avant-garde* but still cannot be considered an error as light as for the *Impressionism/Avant-garde* confusion. Obviously, *Renaissance* and *Avant-garde* are two completely different art movements and confusing them is considered a severe error. The *kappa* value we obtained with this setup is 0.604 which, given the general reference table provided in [**?**] and reported in Table 1, indicates that the agreement lies on the boundary between *moderate* and *good*.

Kappa value	Agreement
< 0.20	Poor
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Good
$0.81 - 1.00$	Very good

Table 1: Kappa value and the corresponding interpretation of agreement.

5. CONCLUSIONS

We have investigated how, during human-robot interaction, tracking the way users explore a structured document concerning works of art helps obtaining implicit feedback. We concentrate on user requests that are differently informed with respect to the predictability of the associated piece of information and check if the proposed measures are consistent with reported user interests. Evaluation is performed with an explicit and an implicit measure to take into account potential interest overestimates users may provide, the topic being works of art. Results show that the experience was positively evaluated and that the agreement on the degree of interest is promising. While in this work we consider a classification based on artistic movements, the procedure applies to other viewpoints, too. Also, a graph based representation allows multiple classifications to co-exist and will be covered in future work.