Toward a Social Attentive Machine

Matei Mancas and Nicolas Riche and Julien Leroy and Bernard Gosselin and Thierry Dutoit
matei.mancas, julien.leroy, nicolas.riche, bernard.gosselin, thierry.dutoit@umons.ac.be
University of Mons - UMONS, TCTS Lab,
31, Boulevard Dolez, B-7000 MONS, Belgium
http://www.numediart.org

Abstract

In this paper, we discuss the design of a new intelligent system capable of selecting the most outstanding user from a group of people in a scene. This ability to select a user to interact with is very important in natural interfaces and in emergency-related applications where several people can ask to communicate simultaneously. The system uses both static and dynamic features such as speed, height and social features (interpersonal distances) which are all acquired using a RGB-Depth camera (Kinect). Those features are combined and a contrast-based approach is able to focus the system attention on a specific user without complex rules. People position with respect to the Kinect sensor and learning of the previous people behavior are also used in a top-down way to influence the decision on the most interesting people. This application is represented by a wall of HAL9000’s eyes that search in the scene who is the most different person then track and focus at him until someone more “different” shows up.

Introduction

This paper deals with a system which is able to choose and to follow the most outstanding person from a group of individuals. The system was initially implemented within the scope of an artistic setup in the context of the NUMEDIART Institute (Numédiart 2011) to automatically select a performer (dancer, music player) and augment his performance by different means like video projections. This project is inspired by the work of (O’Shea 2008), where a multitude of mirrors were moving towards a specific user. To represent an attentive machine, we were inspired by the famous red eye of HAL9000, computer awareness of the Discovery One spaceship in “2001, A Space Odyssey” by Stanley Kubrick. A wall of HAL’s eyes was designed to select the most “outstanding” person in a scene by gazing in his direction (Fig. 1). An idea is to push users to compete to be chosen by the machine, thus to try to differentiate from the other users. The wall of HAL’s eyes could be perceived in a negative way by the users, but instead, they all try to be selected and see all the eyes gazing on them. The system selects a user by bottom-up feature contrast (outstanding = different behavior in space) modulated by top-down information on machine’s personal space (closer to the machine = keen to interact) and learning of the previous user behavior (outstanding = novel in time). The characteristics used are mainly related to 3D motion but also social features like the proxemics (E.T.Hall 1966) of the people present in the scene. Finally, a visual feedback is proposed by embedding in the pupil of the eyes the mirror image of the observed person. The system is decomposed in three parts (Fig. 6). Tracking and feature extraction were conducted by using the Kinect(Microsoft 2010) sensor. The attention mechanism was implemented with Max/Msp(Cycling74 2011). Finally, the visual feedback was realised with OpenFrameworks(Noble 2009).

Figure 1: Wall of Hal’s eyes. The system visual feedback: here all the eyes target only one person.

This paper is organized as follows. Section 2 discusses the proxemics and its uses in the presented attention mechanism. Section 3 introduces the extracted features. Section 4 presents the mechanism of attention that is implemented and details of the bottom-up and top-down components. Section 5 deals with the implementation of the application. Section 6 and 7 are devoted to future developments and conclusion.
Proxemics: a Social Feature for a Social Attentive Machine

People create unconscious territories around them, which define and determine the interactions they can have with other people. Those territories are like invisible bubbles surrounding and keeping them far from each other, unless space has some physical constraints (small room, crowded environment...). E.T.Hall, in his studies about the human behaviour in public space (E.T.Hall 1966), introduced for the first time this concept of interpersonal distances or proxemics. He also proposed the most widespread model of interpersonal spaces: it divides the space around a person in four distinct regions as shown in Fig. 2:

1. Intimate distance (0 to 45 cm): a really close space with high probability of physical contact. It is a distance for touching, whispering or embracing someone. It indicates a close relationship (lovers, children, ...)
2. Personal distance (45cm to 1.2m): distance for interacting with family members or good friends. The other person is at arm’s length, only ritualized touch (handshake, ...) can happen.
3. Social distance (1.2m to 3.5m): distance for more formal or impersonal interactions. It’s the distance people naturally hold when they meet unknown people and establish a communication process with them.
4. Public distance (3.5 to infinity): distance for mass meeting, lecture hall or interactions with important personalities.

Figure 2: Definition of Hall’s personal spaces model. Four areas divide the spaces around people: the intimate space, the personal space, the social space and the public space.

In a way, the measure of these distances is a clue that can tell us how people know each other, we can make the assumption that they are probably part of a group, that they are lovers, etc. People update, control, and adjust their personal spaces continuously: any unauthorised penetration of these spaces will cause a feeling of discomfort. This can lead to an aggressive response of the subject who may feel oppressed by the presence of the intruder. Interpersonal spaces are complex social mechanisms and depend on many parameters: they continuously evolve and adapt to the context. They should be seen as dynamic and elastic territories varying with lot of parameters like: culture, sex, age, gender, size, social position, relationship or physical appearance. Observing the spatial behaviour of people provides higher level information compared to personal low level features (position alone, speed, direction) by introducing special relativity between people.

Features Extraction

The first step for an interactive machine is to extract features from the observed people. For that purpose, we use the Kinect sensor for its ability to extract smooth depth maps in complex illumination conditions. Also libraries as OpenNI (OpenNI 2011) are available to detect human silhouettes and extract anatomical features from skeletons tracking (Fig. 3).

Figure 3: Real-time multi-users tracking, automatic skeleton fitting and personal space modelisation: red = intimate, blue = personal space.

Automatic skeleton fitting

A first issue which has to be solved for an interactive system is the skeleton calibration. The OpenNI library requires people to stay in a specific position (psy-like pose) in order to get a skeleton, which is not very convenient especially when there are several users in public places. Nevertheless, this issue can be solved by saving the parameters of one person (who has to calibrate) and then fit automatically those parameters to any new user. In our case, this solution works well because we mainly use the position of the head and upper body parts.

Behavioural Features

Social Signals  In a previous work (Leroy, Mancas, and Gosselin 2011), we developed tools to study proxemics on ecological scenes. Our objective was to propose a new methodology to study spatial behaviours in natural scenes using 3D cameras. The system can, precisely and in real time, detect users, measure interpersonal distances and displays a virtual view of the personal spaces model proposed by E.T.Hall (Fig. 4). This social spatial information is used here to detect who has the most outstanding spatial behaviour. For this purpose, we compute the mean distance between each person on the scene, based on the position of their 3D centroid. The one with the contrasted mean distance, in terms of euclidian distances, will draw the attention. We also use another social information which is people’s height. Height can give us precious information on the
nature and behaviour of the person present in the scene. We look at who has the most outstanding height to weight the saliency map. A child in a group of adults will draw the attention and the reverse is also true. The scenario, where someone crouches while the others stand up, follows the same rules. It is also an interesting information correlated to the proxemics behaviour and first results, observed with (Leroy, Mancas, and Gosselin 2011), seem to show that the evolution of the size of our personal space can be linked to the size of the people interacting.

Figure 4: Personal Space Augmented Reality Tool (Leroy, Mancas, and Gosselin 2011). Software developed to study proxemics behavior in ecological scene using the Kinect.

Motion Features In addition to the static social features previously described, we also observed 3D motion (speed on the 3 axes). The Kinect sensor makes it easy to detect the position of different users. Based on the estimation of their respective centroid, it is possible to study their movement in a 3D space. The motion speed will be the dynamic feature for the bottom-up part of the process of attention. The slower or faster person, in contrast to the rest of the group, will get a higher saliency.

Who’s Different? Use of Computational Attention

The aim of computational attention is to provide algorithms to automatically predict human attention. The term of attention refers to the whole attentional process that allows one to focus on some stimuli at the expense of others. Human attention is mainly divided into two main influences: a bottom-up and a top-down one. Bottom-up attention uses signal characteristics to find the most salient or outstanding objects. Top-down attention uses a priori knowledge about the scene or task-oriented knowledge in order to modify (inhibit or enhance) the bottom-up saliency. The relationship and the relative importance between bottom-up and top-down attention is complex and it can vary depending on the situations (Mancas 2009). Much research has already addressed the computational visual attention, but these works were devoted almost exclusively to 2D image analysis: up to now little has been done on 3D data (Riche et al. 2011). The arrival on the market of low cost 3D cameras opens up new prospects for developing algorithms for visual attention, which should make the move towards more realistic ambient intelligence systems.

Bottom-up Contrast-Based Attention Mechanism

As stated in (Mancas 2007) and (Mancas et al. 2007) a feature does not attract attention by itself: bright and dark, locally contrasted areas or not, red or blue can equally attract human attention depending on their context. In the same way, motion can be as interesting as the lack of motion depending on the context. The main cue, which involves bottom-up attention, is the contrast and rarity of a feature in a given context. We based our approach on (Mancas et al. 2010). In our case, as the group of people is small, the rarity computation is not relevant, therefore we only use the global contrast. Thus, the first step in this section is to calculate for each feature $i$ ($f_{i,k}$) a contrast between the different users $k$ ($C_{i,k}$):

$$C_{i,k} = \sum_{j=1}^{N} \frac{|f_{j,k} - f_{i,k}|}{N-1}$$  \hspace{1cm} (1)

where $N$ is the number of users. Once all the contrasts for a given feature $C_{i,k}$ between each user and the others have been computed, they are ordered in ascending order $C_{i,k,o}$ with $o = [1 : N]$ from the maximum ($o = 1$) to the minimum ($o = N$). The difference between the two highest values is compared to a threshold $T$ which decides if the contrast is large enough to be taken into account as in Eq. 2.

$$\left\{ \begin{array}{l} \alpha = 0 \text{ if } |C_{i,k,1} - C_{i,k,2}| < T \\ \alpha > 0 \text{ if } |C_{i,k,1} - C_{i,k,2}| \geq T \end{array} \right.$$  \hspace{1cm} (2)

Only the features being the largest and passing this threshold $T$ are merged with different weights (Eq. 3).

$$C_k = \sum_{i=1}^{H} \frac{C_{i,k} + W_i + \alpha}{H}$$  \hspace{1cm} (3)

Where $H$ is the number of features and $\alpha$ is given in Eq. 2. The way the values of the weights $W_i$ is set is described in the next section. This contrast $C_k$ represents the bottom-up saliency for each user $k$. Saliency will be higher for the people exhibiting the most contrasted features (motion, position) within a given frame. For each frame it is possible to highlight the most different behavior: if we take into account the proxemics-related features for example, a person who is isolated compared to the others will attract the system attention. An interesting fact concerning proxemics is that if several people are aligned, it is the individual at the center that will attract the attention which is finally what humans would naturally do (if they would take into account only the interpersonal distances). Also, a person staying still can be salient if all the others move which shows that in our approach motion is not necessary salient by itself (of course most of the time, motion is also rare and salient but in some configurations where static objects are more rare than moving ones, the lack of motion can be surprising).

Top-down Information for Adapative Features Weights

The weights $W_i$ from Eq. 3 can be adapted via the top-down component of attention. Those weights are initially set to be the same for all the 4 features which are used here. Than, the
number of times a feature is contrasted enough for a given user ($\alpha > 0$), a counter is increased. The feature weights will be inversely proportional to their counters: if a feature $i$ is often contrasted, its weight will be lower and lower, while a feature which is rarely contrasted enough will see its weight increased. This mechanisms ensures a higher weight to novel behavior while too repetitive behavior will be penalized. As an example, someone who first will sit down (different height feature compared to the others), the height will have the maximum weight. If this person thinks that a different height is enough to attract the system attention, he will try again, but the more he tries again, the more the height feature weight will decrease as this behavior is no longer surprising. This top-down approach allows the system to learn how much a feature is novel and provide higher weights to the most novel ones. While the bottom-up approach will analyse in each frame which person behaves in a different way compared to the others, the top-down approach will also look for people who behave in a novel (or less repetitive) way from the beginning of the observation.

**Adding Top-Down Social Information**

The weights $W_i$ from Eq. 3 can also be changed via another top-down component. For this purpose, we used the concept of proxemics applied to the system itself. This concept is not only restrained to human-to-human interaction but it can be also applied for human-computer interaction. Some research have already been done on the spatial interaction we can have with computers (Ballendat, Marquardt, and Greenberg 2010) or with virtual worlds (Friedman, Steed, and Slater 2007) and it seems that we often behave in the same way with machines than with humans. Thus, the proposed system was given its own proxemics behaviour. We defined two social territories around our installation, physically represented by the center of the screen (Fig. 5). The contrast equation with the top-down component becomes:

$$C_k = \sum_{i=1}^{H} C_{i,k} = \frac{\sum_{i=1}^{H} C_{i,k} \cdot W_i \cdot \alpha \cdot \beta}{H}$$  

where $\beta = 2$ if the user is located closer to the system and $\beta = 1$ if he is located in the farther space. Fig. 5 shows differences between bottom-up attention alone and top-down modulation adding proxemics information.

**Implementation and User Feedback**

To achieve our application, different modules were used (Fig. 6). Firstly, the graphic look was achieved using OpenFrameworks, a C++ library dedicated to creative coding, bringing together diverse libraries like OpenGL, OpenCV, etc. The first step to create this watching machine was to calibrate the gazes from the eyes on the real world. Like the wall of eyes is not plane but a piece of cylinder, Fig. 7, we design a 3D virtual world, where are drawn the eyes, like an extension of the real world. In that way, we wanted to give the impression that the screen is the frontier between the two worlds. This setup makes it easy to calibrate the installation at the size of the screen and lets the gazes look precisely into the real world since everything is computed on the basis of a single mark that is the Kinect camera. Secondly, to analyse the scene, the OpenNI framework was used, that supports the Kinect camera and allows the design of natural user interfaces. This library is used to realise people detection, segmentation and head-tracking. Detection is for orienting the gazes and provides features to the attention mechanism. Segmentation and head-tracking are part of the visual feedback given back to the users to encourage them to compete to be selected by the machine. First, the tracked person is segmented, then we superimpose the mirror image of the followed person in the pupil of the eyes. Secondly, with the head-tracking possibility, we change dynamically the field of view of the application to create a impression of 3D and the continuity of the real world into the virtual world, Fig. 7. Finally, the implementation of our attention mechanism and the features’ processing was performed using Max / Msp, a platform for real-time signal processing which allows fast prototyping by using visual programing with libraries supported by a large community. To communicate the skeleton information obtained with OpenNI, we developed a small application that sends joints and user information to Max using OSC (OpenSoundControl (OSC 2011)) protocol.

**A Kinect-Based Network**

An issue of the Kinect is the limited interaction space which is very convenient for single user but too limited for more than 4 users. The solution we used to overcome this issue was to calibrate together several Kinects and to fuse the data coming from the two cameras. The implemented solution uses two Kinects but it could be easily generalized to several...
sensors. The number of cameras is limited by the number of USB busses installed on the computer. To solve this first issue, and to augment the number of sensors, we developed an application that works with a master-slave architecture (Fig. 9). The slave instances of the tool realise the capture and tracking then send the users’ position and skeletons to the master with OSC messages. Each slave is provided with a transformation matrix giving its position in the coordinates system of the master. This transformation matrix is obtained by doing a registration step on the overlapping point clouds given by the slave and by the master. To do this, we use spatial filtering to limit the point clouds of each view at the overlapping area then an initial guess of the cameras position is obtain by a manual registration. Finally, we refine the calibration by using the automatic ICP algorithm (Zhang 1994). As we have the position of each cameras, we can fuse all users’ data and get a large interactive scene. Our results are promising. The accuracy of close objects (50 centimeters to 4 meters) is good, while this accuracy is lower and lower when getting far from the sensor due to registration errors. Fig. 8 shows the 3D point cloud coming from two different Kinects after registration.

When the user is swithing from a Kinect to the second one (exterior screens on Fig. 9), his skeleton keeps a precise colour (blue or pink) but it is still there.

**Conclusion and Perspectives**

In this paper, we propose an attention-based people selection system which uses 3D motion features. Moreover, we introduce higher level characteristics based on social behaviour. The proxemics, used in bottom-up and top-down analysis, provides context information on the interaction people could have on the scene. We used this algorithm to make our interactive art installation which aims in augmenting artists performances. Another application is a system which could communicate with one person chosen between several others keen to communicate in parallel which occurs in emergency situations. The system can also be used to study people reaction and interaction in a competing scenario. More generally, a system which is able to choose among several people the one with who it wants to interact in a natural way has plenty of applications for intelligence HCI. Future work will focus, first, to enrich the model of attention with a higher number of features. Tracking on a wider scene provides access to more people, opportunities and social behaviours. It will be interesting to study the dynamics of these different social groups, their creation, their division, their interaction in the same way as we analysed individuals more “isolated”.

**Acknowledgements**

This work was funded by the Numediart Institute (www.numediart.org), Walloon region, Belgium.

**References**


