MMSpace: Kinetically-augmented telepresence for small group-to-group conversations

Kazuhiro Otsuka*
NTT Communication Science Laboratories

Abstract
A novel research prototype, called MMSpace, was developed for realistic social telepresence in small group-to-group conversations. MMSpace consists of kinetic display avatars, which can change pose and position by automatically mirroring the remote user’s head motions. To fully explore its potential beyond previous alternatives, MMSpace has the following novel features. First, it targets symmetric group-to-group telepresence. Second, the kinetic avatars of MMSpace can produce highly accurate, low latency, and silent physical motions, by using 4-Degree-of-Freedom (DoF) direct-drive actuators, and they can express a wide range of natural human behaviors like head gestures and changing attitudes, as well as indicating the focus of attention. Third, MMSpace supports eye contact between every pair of participants, by integrating i) directional visual attention cues indicated by avatar’s kinetic pose change, ii) line-of-sight alignment among the positions of persons, avatars and cameras, and iii) attention-based camera switching, which allows an avatar to always show its owner’s face looking directly toward the person that the avatar’s owner is looking at. The prototype targets the 2 x 2 setting, and subjective evaluations based on group discussions indicate that the kinetic display avatar is superior to static displays in various aspects including gaze-awareness, eye-contact, perception of other nonverbal behaviors, mutual understanding, and sense of telepresence.

Index Terms: H1.2 [Models and Principles]: User/Machine Systems—Human factors; H4.3 [Information Systems Applications]: Communications Applications—Computer conferencing, teleconferencing, and videoconferencing;

1 Introduction
Face-to-face conversation is the most basic form of human communication used for conveying/sharing information, understanding others’ intentions/emotions, and making decisions. Social telepresence is needed in order to make communications between parties in spatially separated places as natural as possible, i.e., as if they are in real face-to-face settings. In recent years, kinetic displays have been gaining attention as prospective elements of embodied avatars for this purpose [37, 1, 34, 33]. In this paper, the term kinetic display avatars, or simply kinetic avatars, refers to a kind of embodied avatar of a remote person that is realized as a flat panel display and that can change its pose and/or position by automatically mirroring remote user’s head motions or manually control inputs by the remote user. This additional physical modality may potentially boost nonverbal exchanges among spatially separated conversation participants and to increase the sense of social telepresence.

The importance of kinetic motions augmenting the displayed images lies in the fact that physical nonverbal behaviors play essential roles in face-to-face conversations [4]. Among the various nonverbal human behaviors, head motions are especially important for expressing visual attention and gestures. Visual attention, also known as gaze, has functions such as watching others, expressing one’s attitudes and intentions, and regulating conversation flow [12]. Head gestures are used to express one’s emotions/attitude, turn-yielding/taking cues, and a back-channel response to signal interest in a speaker [15]. Lack of such nonverbal exchanges results in insufficient reality and the inability to achieve effective telecommunication [32].

A number of pioneering studies have revealed the potential of kinetic display avatars for enhancing the perception of gaze awareness [34, 33, 22, 21], emotion, gestures, posture of people in conversation [33, 21], and the sense of presence of remote participants [21]. However, previous kinetic displays have limitations such as distracting motions triggered by incidental actions of the user, ambiguity in interpreting kinetic motions, audible mechanical noise, and more crucially, lack of eye contact [34]. As such, the full potential of kinetic displays has not been fully realized, and applications have been limited to, e.g., asymmetric hub-and-satellite-type meetings.

This paper aims to explore the potential of kinetic display avatars for social telepresence that approaches the realism of face-to-face settings, the gold standard, and describes a new research platform called MMSpace. The novel features of MMSpace are threefold. First, MMSpace targets symmetric group-to-group (or multi-to-multi) communications, i.e., a mixture of face-to-face and remote communications. Here, symmetric means that every location has the same setting and all participants can equally engage in conversation, in contrast to asymmetric, where remote people use different interface devices and/or environments.

Second, MMSpace aims to produce highly accurate, low latency, silent kinetic motions and incorporates 4-Degrees-of-Freedom (DoF) precision-machine-grade direct-drive actuators that control the pose and position of the display, e.g., a square semi-transparent projector screen panel, which shows the face and shoulder image of the remote participant. We hypothesize that many of the limitations of existing kinetic display avatars are due to immature implementations of mechanical and control systems and that highly accurate kinetic reproduction of human physical motions can provide informative nonverbal cues that will yield effective communication between separated parties and boost the sense of presence without distracting users.

Third, and most importantly, MMSpace aims to offer the sense of eye contact between every pair of participants engaging in multi-to-multi telepresence. Eye contact, also known as mutual gaze, is a key requirement for effective telepresence [27]. Although some of the existing kinetic avatars may be able to support a rough approximation of eye contact, no past study has evaluated the level of success and/or failure of eye contact in their user studies; hence, the eye contact problem remains unclear and unsolved. MMSpace focuses on this issue and implements a mechanism that is expected to make eye contact possible in multi-to-multi situations. The principle is that an avatar always shows its owner’s face looking directly toward the person that the avatar’s owner is looking at, and it is embodied by the multimodal visual attention cues yielded by combining kinetics, optics, and imagery, that can be summarized as i) kinetic changes of display pose, ii) line-of-sight alignment,
and iii) attention-based camera switching. Kinetic cues help the participants understand who is looking at whom, from the facing direction of the avatar displays. Line-of-sight alignment needs the visual parallax (caused by separation between the avatar’s eye on the display and avatar’s camera) to be minimized, and the image of the participants at avatar’s eye position to be captured, as if the avatar were actually looking at the person. To do this, MMSpace sets cameras behind the avatar’s screen to capture the line-of-sight connecting the intended person’s eye and the avatar’s eye on the display. The user’s face image is captured by the avatar’s camera assigned to him/her, and this image is shown on the user’s avatar display. Camera switching selects one of the camera images on the basis of the user’s visual attention, so that the user’s straight-looking-face is provided to the person whom the user is attending. Because the gaze target changes over time, dynamic camera switching is necessary to correctly establish eye contact for multi-to-multi telepresence.

Another distinct feature of MMSpace is that it uses semi-transparent flat panel screens as the avatar’s display (=face), onto which life-size faces of the remote people are projected. The merits of a transparent screen are i) a boosted sense of telepresence: the image of the remote person overlays the actual room background and provides the impression that the other person appears to float in the air just in front of the viewer; ii) cameras are set behind the screen to see through the target person’s image; iii) the panel edge is visible and provides kinetic motion cues to viewers; and iv) the light weight of the screen supports excellent kinetic performance.

Initial experiments in the small group setting (2 × 2) have confirmed that the current kinetic implementation achieves high accuracy (<0.3 mm in translation and <0.2 deg. in rotation) with a very short time lag (<100 ms), shorter than that of the video channel (∼150 ms). Subjective evaluations based on group discussions confirmed that the kinetic avatars are superior to static display avatars in various aspects including gaze-awareness and eye-contact, perception of other nonverbal behaviors, mutual understanding, and sense of telepresence. In summary, the contribution of MMSpace is that it enhances the potential of kinetic display avatars for social telepresence through its higher kinetic performance and mutual eye contact. The author believes that the findings presented in this paper will contribute to better designs for telepresence systems and robots in the future.

This paper is organized as follows. Section 2 describes related work. Section 3 details the proposed system. Section 4 describes the experiments and results. Section 5 discusses the results. Our conclusions are presented in Section 6.

2 Related Works

This section briefly reviews the past kinetic avatars and then clarifies the current innovation and its relation with past studies. Table 1 compares some aspects of previous kinetic display avatars.

| Kinetic display avatars | An early prototype was Porta-Person [37], a portable audiovisual terminal with a small 1-DoF swiveling display. It was intended as a video surrogate for a remote user participating in multiparty meetings. The remote user clicks on a point in the panorama screen of the remote terminal to rotate the kinetic display towards the target. Currently, this type of telepresence terminal can be purchased in the market, e.g., KUBI [29]. RoCo [5] is a computer terminal with an LCD screen supported by a 5-DoF neck; it can present various physical postures. It was found that the physical posture of the screen can influence the cognition of users; e.g., users tend to mirror the posture of the robot screen. MeBot [1] is equipped with a small 4.13” display mounted on a 3-DoF neck (head-pan, head-tilt, and neck-forward); pose is automatically controlled by using face tracking on the remote user. MeBot was used to explore the expressivity of kinetic avatars, and it was indicated that users found kinetic avatars more engaging and likable than static ones.

Sirkin et al. targeted an asymmetric one-to-multi setting (they called hub-and-satellite) and implemented a 1-DoF swiveling display controlled either manually or automatically by face tracking [34]. Their experiment suggested that swiveling displays can enhance directional attention cues. A trade-off between manual and automatic control was found such that manual control could more clearly express the intention of the remote user, but entailed a delay in response. In contrast, automatic control could reduce the cognitive load in operation, but raised more problems with ambiguity in motion interpretation and distracting unintentional motions. Later, Sirkin et al. developed a 3-DoF avatar (they called it kinetic proxy) that was manually controlled by the remote user [33]. They studied the impression of viewers who watched short video clips including interactions involving a single proxy and two humans. They stated that the combination of both on-screen and in-space motions can enhance the viewer’s understanding of certain types of gesture, e.g., visual attention. MMSpace addresses and tries to overcome the limitations of these previous systems, while enhancing the merits of kinetic displays.

For visualizing remote multiparty conversations conducted offline, the author’s group proposed a 2-DoF (pan and tilt) kinetic display, MM-Space [23, 22] and a 4-DoF (+2-D horizontal translation) kinetic display called MM+Space [21]. Their experiments indicated that kinetic displays can enhance the perceptions of gaze direction, emotion, gestures, and posture of people in conversation and can boost the sense of presence of remote participants. However, their users felt that the physical head motions were sometime unnaturally overemphasized, distracting, and ambiguous. Comparing the 2-DoF and 4-DoF approaches, it was found that 4-DoF displays outperformed 2-DoF ones at expressing posture and the sense of telepresence [21]. MMSpace improves the mechanics of MM+Space; e.g., geared motors for 2-DoF rotations are replaced with silent direct-drive motors, for richer kinetic expressibility and less mechanical noise.

Table 1: Comparison of kinetic display avatars

| Porta-Person(2007)[37] | one-to-multi | no | no | 1 | manual | not addressed |
| MeBot(2010)[1] | one-to-one | no | no | 3 | auto | one-way/not evaluated |
| Sirkin et al. (2011)[34] | one-to-multi | no | no | 1 | auto/manual | no |
| Sirkin et al. (2012)[33] | one-to-multi | no | no | 3 | manual | not addressed |
| MMSpace(2011) [23, 22] | 4-party playback | — | yes | 2 | auto | — |
| MM+Space(2013)[21] | 4-party playback | — | yes | 4 | auto | — |
| MMSpace (this paper) | multi-to-multi | yes | yes | 4 | auto | yes |
| Shader lamp avatar (2009) [14] | one-to-multi | no | humanoid | 2 | auto | approximate/not evaluated |
Note this paper distinguishes kinetic display avatars from humanoid-head kinetic avatars, which aim to simulate the physical presence of humans, such as the Shader Lamp Avatar [14, 30] and others [18, 16, 3]. This is because the key to kinetic displays is the integrated effect of kinetic motions and image motions; i.e., even a display with a simple geometric shape can trigger a strong sensation of a lively human presence. The merit of the humanoid head is that it can offer more accurate gaze cues compared with a flat display avatar. The drawback is, however, that making personalized head/face shaped displays requires rather complicated processes and runs the risk of entering the uncanny valley [17].

Figure 1: MMSpace. (a) and (b) are the corresponding kinetic avatars of persons (d) and (c). (a) and (c) are in room 1, and (b) and (d) are in room 2. (e), (f), and (g) show sample conversation scenes.

Eye contact  Eye contact or mutual gaze has always been recognized as a key requirement for effective visual communications [27]. Previous studies have not fully explored this issue, and hence, it remains unresolved. The inability to establish eye contact is rooted in the visual parallax created when the camera position is offset from the eye position displayed on the screen. When a user looks at a face shown on the display, the camera captures the user’s face, but the eyes are not directed towards the camera. Thus, a viewer at a different location sees the remote person as exhibiting an averted gaze. This causes a strange sensation and hampers natural telecommunication, because human vision is very sensitive to visual parallax [7]. Most existing kinetic displays suffer from poor eye contact because the camera is fixed atop the display [37, 1, 34, 33, 14]1. Even in this setting, if the display is relatively small so that distance between the camera and displayed face is short, i.e. the visual parallax is small, approximate eye contact could be possible. However, none of the past studies on kinetic display avatars reported on the level of success or failure of eye contact.

Furthermore, as an interface for the satellite user, a single screen is often used to display the images of multiple persons and a camera may be located at the center of the display [34]. When the speaker looks at a person located on the side of the screen, the speaker’s face appears to have “turned away” on the screen in the other rooms. Hereafter, we refer to this as the turn-away effect. In this case, the participants have difficulty in knowing who is looking at whom and find it impossible to establish eye contact.

For realizing better eye-contact in multi-to-multi telepresence, a more complex situation than those assumed by previous avatars has to be taken into account. Here, MMSpace embodies an integrated kinetics-optics-imagery system that exchanges accurate and useful visual attention cues among participants. For minimizing the visual parallax, i.e. ensuring line-of-sight alignment in the eye contact mechanism, MMSpace takes an approach that places cameras behind a semi-transparent screen [27]; such an arrangement has been used in [10, 20, 19]. Although a well-known idea, this study is the first to implement it in a kinetic display avatar.

3 System and Implementation
This section presents the system and implementation of MMSpace as a proof-of-concept of kinetic display-based telepresence for small group-to-group (e.g. 2 × 2) conversations.

3.1 Kinetic display avatars and configurations
To support mutual eye-contact, MMSpace is designed to provide multimodal visual attention cues by integrating kinetics, optics, and imagery. The following paragraphs detail each element.

Spatial configuration  Fig. 2 illustrates the spatial configuration of MMSpace for the case of symmetric 2×2 telepresence. In room 1, two people, person 1 and person 2 (hereafter denoted P1 and P2), and two kinetic avatars, A1 and A2, of the respective remote participants, P3 and P4 in room 2, are situated around a round table. Spatial consistency across the rooms is a prerequisite for delivering and sharing visual attention cues among people situated in separate rooms.

1The creators of the Shader Lamp Avatar suggested that the camera could be embedded in the eye of the humanoid head (avatar), but this idea was not implemented.
Kinetic display avatar  Fig. 3 shows a frontal view of an avatar screen. The screen is a highly transparent acrylic panel but includes a diffusive material that catches the projector’s output and makes it visible to the viewer. It offers a wide viewing angle (e.g., 160 deg.) and does not have a critical angle, unlike holographic optical elements. Each screen has its own LCD projector behind it. Fig. 4(a) shows the back of the avatar with its actuators. Fig. 4(b) illustrates the details of the actuators, which consist of an XY-stage, a rotation-stage, and a gonio-stage, from bottom to top. The XY-stage generates horizontal translations along the X axis (left-right) and Y axis (forward-backward). The rotation-stage generates rotational motion around a vertical axis, which corresponds to the head turning or shaking. The gonio-stage generates rotations around the horizontal axis, which corresponds to the head nodding. All these motor stages use direct drive motors, which are silent, unlike geared DC motors used in most previous kinetic avatars, which emit noticeable audible noise. The kinetic display avatar can express a wide range of human head motions, including turning for orienting one’s visual attention towards others.

The high transparency of the avatar’s screen is expected to contribute the sense of presence of remote people; i.e., the remote persons appear to be in the same room. A disadvantage is that visibility varies with the background color and texture, which can be seen through the panel. To alleviate this issue, for simplicity, black wall screens were installed behind each panel. In addition, to increase the sense of presence, the avatar panels display only face and shoulder images. This is done by removal of the background in the image. For simplicity, the current version of MMSpace places a black screen behind each person.

Camera configuration and camera switching  As shown in Fig. 2 and Fig. 5, cameras are placed behind each avatar’s screen panel, to capture the images of each person in the room, from the avatar’s point of view. More specifically, the camera is aligned along the person’s line of sight, which starts at the person’s eye and penetrates the avatar’s eyes on the panel (the face image is always projected at a fixed position on the panel). Hereafter, eye refers to the middle point of both eyes. Because of the high transparency of the screen panel, the cameras behind it can capture clear face images. To eliminate the reflection of the projected image, the cameras and the projector have crossed polarization filters. Each avatar has multiple cameras pointed at each person. For example, the images of $P_i$ ($i \in \{1, 2\}$) in room 1 are captured by two cameras, $C_{1 \rightarrow j}$ and $C_{2 \rightarrow j}$, which correspond to the viewpoints of avatars $A_3$ and $A_4$, respectively. One camera image of the person is selected and projected on the screen of avatar $A_i$ in room 2 by using the projector behind the avatar.

The camera selection/switching is based on the direction of visual attention. Here, MMSpace uses the principle that an avatar always shows its owner’s face looking directly toward the person that the avatar’s owner is looking at. For example, when person $P_1$ in room 1 looks straight at avatar $A_3$’s face on the screen, avatar camera, $C_{3 \rightarrow 1}$, captures an image that is close to that of the person directly gazing at the camera. This camera image is selected and displayed on avatar $A_1$ in another room. In room 2, if person $P_1$ looks at avatar $P_1$, $P_3$ can see $P_1$’s image looking straight at him/her. At the same time, avatar $A_1$’s camera, $C_{1 \rightarrow 3}$, captures an image of $P_3$’s face looking straight at avatar $A_1$, and displays it on avatar $A_3$ of $P_3$. Accordingly, $P_1$ should realize, from avatar $A_3$’s image, that he/she is being looked at by $P_3$. Through this process, mutual eye contact between two people, $P_1$ and $P_3$, is established. In the multi-to-multi setting, because there is more than one gaze target and the gaze target changes over time, dynamic camera switching is used to select one of the camera images per person, on the basis of the direction of visual attention of each person, for correctly establishing eye contact between every pair in the conversation.

Thanks to the kinetics built into the avatar, the panel pose is linked to the actual head pose; when a person turns his/her face towards another, focus of attention is expressed by the panel’s pose dynamically changing; it faces the other person. This combination of a kinetic display and attention-based camera switching is expected to create the sensation of eye contact between every pair of conversation participants. Fig. 6 shows sample eye contact scenes, including front-to-front ($P_1$-$P_2$), side-to-side ($P_3$-$P_4$), and side-by-side ($P_3$-$P_2$) situations in the same room. In addition, we can expect that other participants can perceive the eye-contact situation from the panel poses.

3CAD data of the motor stages were provided by Aerotech Inc. [2].
sor position and pose motion capture system, Polhemus Fastrak, captured the 6-DoF sensed over the movable ranges of the panel and the person, which however, preliminary experiments confirmed that eye contact is as-
home position, as indicated in Fig. 2, the line-of-sight linking the

Processing of head pose and position A magnetic-based
sensor position and pose \( [x, y, z, a, e, r]^T \) at 60 [Hz] in global (table-centered) coordinates. Here, \( [x, y, z]^T \) denotes the 3-D position, and \( [a, e, r]^T \) denotes the 3-DoF rotation angles of azimuth, elevation, and roll. The sensor was attached to a headset microphone at the right temple of the user’s head. First, a 1-Euro filter [6] was applied to remove measurement noise. Then, the data was converted into the head center coordinate of each person, as shown in Fig. 2. Next, the face center position, which is defined here as the middle point between the eyes, was calculated from the head-center position. The face-center position projected on the 2-D image plane of each camera was calculated in order to crop the face region in the image.

**Estimation of visual attention** Camera switching was used to select the image to be displayed. This was achieved by estimating the gaze target of each person from the head pose. Head pose is considered to be a reasonable indicator of eye gaze [35, 24], because a person tends to focus his/her attention on the person of interest by centering that person in his/her visual field, which results in a rotation of the head and/or torso. Here, **MMSpace** employs a simple gaze estimation scheme based on a Gaussian distribution-based likelihood. When person \( P_i \) looks at target person \( P_j \) or avatar \( A_j \), the likelihood function, \( L_{i,j} \), is defined as

\[
L_{i,j} = N(a | \bar{a}_{i,j} - v_i, c_i, \sigma^2_j),
\]

where \( N(a | \mu, \sigma^2) \) denotes the Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \), and \( \bar{a}_{i,j} \) denotes the azimuth angle from the face center of person \( P_i \) to the face center of \( P_j \) or \( A_j \). Here, \( v_i \) is the velocity of the azimuth change, and \( c_i \) is a coefficient. This velocity term makes the gaze estimation sensitive to subtle head turns.

On the basis of the likelihood values, the gaze target is estimated using a rule that the gaze target changes to \( k \), if and only if all of the following conditions are satisfied: i) target \( k \) has a higher likelihood \( L_{i,k} \) than those of the others \( L_{i,j} \); ii) the likelihood ratio \( (L_{i,k}/L_{i,j}) \) is above a threshold, where \( l \) denotes the previous gaze target; and iii) the elapsed time since the last change in gaze target exceeds a threshold.

In addition, an attention status, looking or averted, is estimated from the elevation angle of the head pose and the relative positions of the person and the potential gaze targets. To do this, a Gaussian-based likelihood function is defined and used in a similar way to the gaze target estimation.

**Camera switching** On the basis of the estimated gaze target and attention status, one of two cameras is selected for each person, e.g., when person \( P_i \)‘s gaze target is \( A_k \), the camera of \( A_k \) facing \( P_i \), i.e., \( C_{k,i} \), is selected, with the following exceptions: i) when the person’s attention status is averted, no camera switching occurs; ii) when \( P_i \)’s gaze target is another person in the same room, the camera facing \( P_i \) (e.g., \( C_{5,1} \) for \( P_1 \)) is selected.

**Actuator control** The head coordinates, \( x \) and \( y \), are used to control the \( X \) and \( Y \) axes of the \( X \) and \( Y \) stage, respectively. Head poses \( a \) and \( e \) are respectively used to control the rotation axis, \( R \), of the rotation stage and the gonio axis, \( G \), of the gonio stage. Because the actuators have limited travel ranges, the head position and pose are mapped to bounded variables, \( x' \), \( y' \), \( a' \), \( e' \), using a sigmoid function such as

\[
x' = TX \cdot \text{sig}(sx) / TX,
\]

where \( \text{sig}(x) := 2/(1 + \exp(-2x)) - 1 \). Here, \( sx \) is a scale parameter; \( sx = 1 \) means that the speed along the \( X \) axis becomes the same as the original speed at the home position \( (x = 0) \), e.g., \( sx = sy = sr = 1 \) and \( sG = 0.4 \). \( TX \) denotes the maximum limit position of the \( X \) axis (e.g., \( TX = 45 \) mm, \( TG = 8 \) deg, \( TR = 35 \) deg if the panel turns outward, and \( TR = 20 \) deg if the panel turns inward). This asymmetric mapping function for the rotation axis enables the following panel behavior. When an avatar turns to a person to make eye contact, the avatar turns its face more to the person, compared with the case that the avatar turns towards another avatar.

Next, a PD (proportional-derivative) controller is used to control each axis of the 4-DoF actuators. The goal of the control is
to ensure that the avatar precisely follows the actual human head movements, which are represented as time series data of the target position \([x', y', d', \alpha']\). Here, the error between the current position and the target position of each axis is minimized, and as a control variable/command, the velocity component along the axis, \(u_t\), is fed to the motor. For each time step \(t\), the velocity is calculated as \(u_t = K_p \cdot e_t + K_d \cdot \frac{d(e_t)}{dt}\), where \(e_t = X_t - x'_t\), and \(X_t\) denotes the current position of the axis, which is obtained from the encoder feedback of each axis, and \(x'_t\) denotes the target position. \(e_t\) represents the error between the current position of the stage, \(X_t\), and the target position, \(x'_t\). \(K_p\) and \(K_d\) are coefficients for the \(P\) term and \(D\) term of the PD control, respectively.

**Projection mapping** Finally, the face image of each person is projected onto each avatar’s screen panel. Because the avatar panel dynamically changes its position and pose, dynamic projection mapping is required to continuously project a skew-free image. A perspective projection matrix is calculated from the relative positions of the projector and the avatar’s position/pose. Here, to compensate for the time lag between the measurement of the axis positions and the actual projection by the projector, the predicted axis position/pose is used to calculate the projection matrix. For more details, see [22, 21].

### 4 Experiments and evaluations

Experiments and evaluation were conducted on a prototype MMSpace system, as follows.

#### 4.1 Hardware

Each digital camera was a Point Grey Research Grasshopper GS2-FW-14S5C-C (XGA@30fps). Each projector was an Epson EB-1965 (5000 lm, XGA resolution). The motion capture device was a Polhemus FASTRAK (60 Hz/person). A Yamaha YVC-1000 teleconference terminal was used. The XY, goniostage, and rotation stages were Aerotech units (ANT130-110-X/Y, ANT-20G-90, and ANT130-180-R) [2]. The travel range and max speed of the XY stage were 110 × 110 mm and 350 mm/s on each axis. The goniostage covered ± 10 deg, and the maximum speed was 150 deg/s. The rotation stage had a 180 deg range, and its maximum speed was 120 deg/s. The screen panel was a Prodisplay Clearview Acrylic; its viewing angle was 160 deg, and transparency was 97%. The screen size was 415 × 415 × 3 mm. The PCs were equipped with an Intel Core i7-3960X@3.3GHz, and 2 Dual GPU cards (NVIDIA GeForceGTX690). One GPU core was assigned to image processing, and two GPU cores were assigned to MJPEG encoding. The OS was MS Windows 7 64-bit.

#### 4.2 System performance

**Audio-visual latency** The latency of video was approximately 150 ms, which was measured as the elapsed time from when the camera’s shutter opened until actual projection started. Audio latency was approximately 35 ms (5 [msec] in the mic. part and 30 [msec] in the speaker part). To harmonize audio-visual latency to 150 ms, extra latency was added to the audio signal by a processor (Yamaha SPX2000).
Table 2: Accuracy of kinetic motion in terms of mean absolute error (MAE) and standard deviation (STD) between the encoder feedbacks (actual positions) and target positions, and time lag. Units are [mm] for the X and Y axes, and [deg] for the gonio and rotation axes. The left column shows the original MAE, and the middle column shows the MAE of time-shifted positions, with the calculated time lag shown in the right most column.

<table>
<thead>
<tr>
<th></th>
<th>MAE(STD)</th>
<th>MAE with time-lag</th>
<th>Time-lag[ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0.590(1.883)</td>
<td>0.211(1.600)</td>
<td>77.2</td>
</tr>
<tr>
<td>Y</td>
<td>0.561(1.562)</td>
<td>0.266(1.169)</td>
<td>81.0</td>
</tr>
<tr>
<td>Gonio.</td>
<td>0.140(0.192)</td>
<td>0.082(0.115)</td>
<td>56.6</td>
</tr>
<tr>
<td>Rotation</td>
<td>0.641(1.427)</td>
<td>0.136(0.349)</td>
<td>93.2</td>
</tr>
</tbody>
</table>

4.3 Experiment 1: Eye-contact perception

An experiment was conducted to check if MMSpace could actually offer the eye contact sensation as expected. Eight female subjects participated. A pair of subjects were seated in P1 and P2’s positions in room 1, and two experimenters sat in P3 and P4’s positions in room 2. A male experimenter in P3’s seat turned his face to A1, A2, or P1 and kept looking for a while. Meanwhile, each subject answered whether she felt eye contact with him by pushing a handheld button. This trial was repeated several times per target under various combinations of conditions, such as with/without kinetic motion and with/without camera switching.

Table 3 summarizes the percentages of the reported eye-contact sensations. Without camera switching, due to the turn-away effect, the person on the side P1 never felt eye contact with P2, denoted as †. With camera switching, however, P1 perceived the eye contact of P2. However, with camera switching, a strong Mona Lisa effect was observed, as indicated by ‡ in Table 3. The Mona Lisa effect [9] is an inevitable effect of viewing flat panel displays; viewers feel eye contact sensations from frontal face images over a wide range of relative angle to the display, not just normal to the display. Here, when P2 looked at the person in front P2’s avatar A2, the person on the side P1 also felt eye contact. A comparison of the with and without kinetic motion conditions suggests that kinetic motion can decrease the Mona Lisa effect (93.8%→63.6%) of the person on the side P1. But this effect was vague and indefinite for the person in front P1 when P2 looked at person on the side A1.

In summary, this experiment confirmed that i) camera switching is necessary to enable eye contact between every pair in the conversation, ii) MMSpace actually offers the eye contact sensation, but it also creates a stronger than expected one due to the Mona Lisa effect, and iii) the kinetic avatar, which faces its display towards the gaze target, potentially offers more correct eye contact than the static display avatar can provide.

4.4 Experiment 2: Group conversation experiments

Experiments based on group discussions were conducted to characterize the impact of the physical motions possible with MMSpace. To do this, motion conditions M and static conditions S were compared. The M conditions used the MMSpace described in this paper. The S condition used a version of MMSpace that did not use kinetic motions, but all other parameters/factors were kept the same as in M conditions, including the spatial configuration and camera switching rule.

Subjects Sixteen paid subjects (hereafter called the participants), who had never experienced MMSpace, participated in this experiment. They were all females in their 20’s ~ 40’s. All participants met for the first time on the day of the experiments. They were separated into four 4-person discussion groups.
Procedure  For each group, a series of discussions were held as follows. The discussion type was consensus making. The participants were instructed to have a discussion on a given topic and try to reach a conclusion as a group within 10 minutes. After each discussion, the participants were asked to fill in a questionnaire. For each group, the experiment began with a self-introduction session in a face-to-face (denotes f2f) setting that used the same round table and seating arrangement as used in MMSpace. The first discussion was also held in the f2f setting, with the topic “Decide three one-day attractions/ spots for a tourist who is visiting Japan for the first time”. Next, four MMSpace-based discussions were held; two were under the motion condition M, and the other two were under the static condition S, in alternating order. Two groups experienced the M condition first, the other groups experienced the S condition first. Discussion topics for MMSpace sessions were “Are love and marriage the same or different?”, “Arrange a travel plan for your group”, “Decide on three Japanese dishes for foreign guests”, and “Which do you desire in a marriage partner, personality or income?”. The order of topics was balanced among the groups.

After all discussions ended, participants were asked to fill in a free-description form as to the pros and cons of MMSpace, followed by a 15-min-long free discussion in the f2f setting with the same topic.

Questionnaire  The questionnaire consisted of 22 common questions for all conditions, 11 additional questions for both M and S, and five questions only for M. The questions can be categorized into general, gaze, addressing, understanding, behavior recognition, sense of presence, behavioral contagion, image/audio quality, and avatar’s physical motion (only for M). Table 4 lists excerpts of the questions. For each question item, a 7-point Likert scale (1: strongly disagree to 7: strongly agree) was used to measure the subjective impressions of the participants.

Results and Discussion  Table 4 summarizes the subjective impressions on the excerpted questions in the questionnaire. For each question item and condition (M and S), the mean and standard deviation were calculated from all sessions and participants. To assess the difference in score between the M and S conditions, the p-value and significance level shown in Table 4 were calculated using a paired t-test. Here, each observation pair consisted of the scores for M and S for each participant, which were averaged over two sessions per condition.

Regarding the responses to the general questions, those for Q1~Q3 suggested that participants easily and naturally communicated with remote participants in both conditions. The responses to Q4 indicated that the kinetic displays put less cognitive load on the participants than the static displays. Q5 and Q6 were gaze-related questions. Their responses indicated that the condition was superior to the S condition for understanding the gaze direction (Q5) and for eye contact perception (Q6). There was no statistical significance for Q7 (who is talking to whom). The responses to Q8 indicated that the M condition helped the participants better understand the reaction of the remote persons. Although there was no statistical significance found in the responses to Q9 (understanding others), those for Q10 indicated that M outperformed S in the sense that participants felt they were better understood by the remote person. Combining the results for Q9 and Q10 indicates that M offers better mutual understanding than S does with remote people. Q11 to Q13 deal with the understanding of nonverbal behaviors, including facial expressions, gestures, and pose/attitude. Here, the responses to all of them indicated that M was superior to S. This indicated that adding physical motions to the display can more clearly express people’s bodily behavior. Interestingly, Q11 indicated that kinetic displays can enhance the perception of facial expressions, although both conditions provided the same facial image on the display. Next, Q14 to Q16 asked about the sense of telepresence. The results showed that M yielded a stronger sense of telepresence, Q17 and Q18 related to behavioral contagion, in particular mirroring and synchrony. M gave participants stronger sensations than S did. Q19 and Q20 related to image perception. Although participants could easily watch the displayed images in both conditions (Q19), their responses to Q20 indicated that M provided a better impression of camera switching than S did.

In addition, the scores from the f2f session are shown in the f2f column of Table 4 for reference, because the score was obtained before participants experienced MMSpace, and a direct comparison of the scores for M and S is thus impossible.

Furthermore, the items in the questionnaire specific to the M condition with kinetic motions detail the participant’s impression of distraction (2.78), naturalness (4.94), ease of interpretation (5.44), and impression of exaggeration (2.94); the average scores are parenthesized. The scores indicate that MMSpace successfully ameliorates the deficits of the previous kinetic avatars, such as distracting motions and ambiguity in motion interpretation.

5 Discussion and Future Work
In addition to the evaluation in 4.4, the impacts of kinetic display avatars were also assessed using a free description form and a reflection discussion. Most of the participants mentioned that conversations using kinetic avatars were totally different experiences than those using static avatars, and they preferred the kinetic avatars. Also, they gave insightful comments such as “For better communications, I realized when speaking, the important things are looking at the partner’s eye, listening and expressing yourself with your whole body. The panel’s motions were proportional to such actions”, and “The panel taught me the importance of nodding to the partner and looking into her eyes”.

In addition to the subjective evaluation in this paper, an objective behavior analysis would be important for understanding the range of communications possible with MMSpace. A comparison with face-to-face data captured using the same group and same seating arrangement would reveal how close MMSpace meetings can
Table 4: Comparison of motion and static conditions. The mean (standard deviation) values on a 7-Likert scale (7: Strongly agree to 1: Strongly disagree) for each question in the post-questionnaire are shown. The significance levels and p-values were calculated by paired t-Test. In the sig. column, the significance level is indicated by * (p < 0.05), **(p < 0.01), ***(*p < 0.001), and ****(p < 0.0001), where ns denotes no significance.

<table>
<thead>
<tr>
<th># Questions</th>
<th>Motion</th>
<th>Static</th>
<th>p-value</th>
<th>sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Easy to communicate</td>
<td>5.63(1.34)</td>
<td>5.22(2.31)</td>
<td>0.165</td>
<td>ns</td>
</tr>
<tr>
<td>2 Naturalness of communication</td>
<td>5.53(1.50)</td>
<td>5.06(1.41)</td>
<td>0.105</td>
<td>ns</td>
</tr>
<tr>
<td>3 Feel unity as group</td>
<td>5.72(1.55)</td>
<td>5.63(1.50)</td>
<td>0.738</td>
<td>ns</td>
</tr>
<tr>
<td>4 Feel fatigued</td>
<td>2.63(1.50)</td>
<td>3.22(1.50)</td>
<td>0.042</td>
<td>*</td>
</tr>
<tr>
<td>5 Easy to know whom remote people looked at</td>
<td>5.69(1.28)</td>
<td>4.94(1.34)</td>
<td>0.016</td>
<td>*</td>
</tr>
<tr>
<td>6 Feel eye contact with remote people</td>
<td>5.88(1.10)</td>
<td>5.34(1.29)</td>
<td>0.042</td>
<td>*</td>
</tr>
<tr>
<td>7 Easy to know whom remote people addressed to</td>
<td>5.41(1.52)</td>
<td>4.91(1.17)</td>
<td>0.056</td>
<td>ns</td>
</tr>
<tr>
<td>8 Well understood the reaction of remote people</td>
<td>5.63(1.18)</td>
<td>5.00(1.32)</td>
<td>0.013</td>
<td>*</td>
</tr>
<tr>
<td>9 Well understood what remote people meant to say</td>
<td>5.75(1.19)</td>
<td>5.47(1.14)</td>
<td>0.095</td>
<td>ns</td>
</tr>
<tr>
<td>10 Feel remote persons well understood what you meant to say</td>
<td>5.66(1.15)</td>
<td>4.97(1.20)</td>
<td>1.815 × 10^-3</td>
<td>**</td>
</tr>
<tr>
<td>11 Well recognized facial expression</td>
<td>6.03(0.93)</td>
<td>5.50(1.02)</td>
<td>0.016</td>
<td>*</td>
</tr>
<tr>
<td>12 Well recognized gestures</td>
<td>5.63(1.36)</td>
<td>4.22(1.58)</td>
<td>1.095 × 10^-4</td>
<td>***</td>
</tr>
<tr>
<td>13 Well recognized pose and attitude</td>
<td>5.59(1.10)</td>
<td>4.25(1.52)</td>
<td>3.218 × 10^-3</td>
<td>**</td>
</tr>
<tr>
<td>14 Feel remote people were close to you</td>
<td>5.78(1.36)</td>
<td>4.69(1.77)</td>
<td>1.343 × 10^-3</td>
<td>**</td>
</tr>
<tr>
<td>15 Feel familiarity with remote people</td>
<td>5.84(1.11)</td>
<td>4.97(1.33)</td>
<td>3.034 × 10^-3</td>
<td>**</td>
</tr>
<tr>
<td>16 Remote people looked lively</td>
<td>5.59(1.52)</td>
<td>4.56(1.58)</td>
<td>3.746 × 10^-3</td>
<td>**</td>
</tr>
<tr>
<td>17 Unintentionally moved triggered by remote people’s motions</td>
<td>4.69(1.77)</td>
<td>2.91(1.53)</td>
<td>1.199 × 10^-3</td>
<td>**</td>
</tr>
<tr>
<td>18 Unintentionally gazed at where a remote person looked</td>
<td>4.66(1.79)</td>
<td>3.34(1.56)</td>
<td>7.272 × 10^-3</td>
<td>**</td>
</tr>
<tr>
<td>19 Easy to watch remote people’s image</td>
<td>5.61(1.15)</td>
<td>5.00(1.19)</td>
<td>0.058</td>
<td>ns</td>
</tr>
<tr>
<td>20 Camera switching was natural</td>
<td>5.00(1.30)</td>
<td>4.13(1.54)</td>
<td>0.017</td>
<td>*</td>
</tr>
</tbody>
</table>

be to face-to-face ones. An example of such an analysis would include speech length, overlaps, reaction time, turn transitions between rooms, correct addressing/reception, and eye contact. Furthermore, since the users’ impression suggested the potential for the avatar’s kinetic motions to induce behavioral contagion, an analysis of interpersonal synchrony among participants would be an interesting topic.

The current version of MMSpace estimates the gaze direction from the head pose and uses it to switch among the cameras. However, it was observed that many participants tended to face the midpoint of the two avatar panels and use eye movement to alternate between the panels. This eye-gaze only attentive behavior does not trigger camera switching. Moreover, subtle head movements around the midpoint of the two panels triggered rapid fluctuations in switching between cameras. A more accurate gaze estimation is required, such as one using both head pose and eye direction [8] and/or using the structure/context of conversations [24]. Moreover, for optimizing the camera switching algorithm, the effect of camera switching on human perception should be investigated, jointly with the effect of dynamic kinetic motions, because pure gaze-based switching might be too fast to be clearly perceived by conversation partners.

Experiment 1 confirmed a strong Mona Lisa effect, as well as the potential that the pose change of the avatar panel could reduce the effect. Hecht et al. reported that the break point of the Mona Lisa effect was 38 degrees away from the frontal direction, in the case of a flat physical surface [9]. The current MMSpace implementation uses 35 degrees as the maximum rotation angle towards a person on the side, and the sigmoid mapping function suppresses the angle to less than the actual pose. Thus, a simple solution is to tune the mapping function and the maximum angle. Also, the limitation of flat displays in expressing the gaze direction was recently reported in [11]. Because all existing studies on the Mona Lisa effect have targeted only the static situation, it is necessary to investigate the Mona Lisa effect under panel motion, with the goal of improving the eye contact made possible by MMSpace. Also, different display shapes are worth considering, including curved surfaces such as geometrical shapes (e.g. cylinders[13][26] and spheres [25]) and more human-like shapes [18, 16, 14, 30, 3].

In addition, many participants in our experiments mentioned the poor usability of the headset. Personally customized and/or lightweight headsets are needed for long-term use. To avoid this problem, image-based sensing of the person’s head position/pose would be desirable, with careful consideration of the measurement stability and the latency created by image processing.

From the viewpoint of visibility, the semi-transparent panel suffers from unwanted light reflections on both sides. To decrease reflections, the screen panel of the next version of MMSpace will have an anti-reflection (AR) coating. The current panel of the displays is only big enough to show the face above the shoulder and seldom captures hand gestures. Hand gestures are important non-verbal behaviors, and their visualization is strongly recommended. A solution would be to enlarge the panel at the cost of higher mechanical loads. Another approach is to use the table surface as a projector screen, e.g., projecting the hands and/or shadows on the table beneath the panel. Such a table projection would allow for the creation of a shared working surface among remote sites, and projected hand images would be highly effective in passing along gestures like pointing.

The optimum motion design and/or adaptive motion control for kinetic avatars is another prospective research topic. The goal of this study was faithful reproduction of human motion by kinetic avatars. However, the optimum motion control might vary depending on the situation or type of conversation, personal communication skills and styles, interpersonal relationships among participants, and latency with telecommunications. One possible starting point would be to amplify or diminish one’s motion on each axis and investigate the effect on the impression and behaviors of individual users and their interactions.

Finally, the scalability of MMSpace should be considered for wider range of teleconferences, more than the 2 × 2 configuration used in this paper. Theoretically, MMSpace can be extended to more participants over more than two places, provided they all sit around a single round table. However, the number of cameras
needed in total grows rapidly with the number of participants, because each avatar needs to have multiple cameras dedicated to each person in the other site(s), calculated as (the number of participants) × (the number of remote participants for each person), e.g., 18 cameras for 3 × 3, 32 for 4 × 4, and 24 for 2 × 2 × 2. In addition, the head-pose-based attention estimation becomes more unreliable when more people join the conversation. Considering implementation complexity, cost, and gaze accuracy, it is thought that current MMSpace can target only small-scale meetings, and scalability remains an open problem.

6 Conclusions
This paper proposed MMSpace as a way to achieve realistic social telepresence for small group-to-group conversations. It consists of kinetic display avatars that can change the screen pose and position by automatically mirroring the remote user’s head motions. To fully explore the potential of kinetic display avatars, MMSpace has the following novel features. First, it is intended for symmetric group-to-group telepresence. Second, its kinetic avatars can produce highly accurate, low latency, and silent physical motions, by using 4-Degree-of-Freedom (DoF) direct-drive actuators, which can express a wide range of natural human behaviors like head gestures and changing attitudes, as well as indicating the focus of attention. Third, MMSpace supports eye contact between every pair of participants. The prototype targets a 2 × 2 setting, and subjective evaluations based on group discussions indicated that the kinetic display avatar is superior to the static version in various aspects including gaze-awareness, eye-contact, perception of other nonverbal behaviors, mutual understanding, and sense of telepresence.

References