

リアルタイムキャプチャを用いた裸眼3Dライブシステムの開発と評価

柳田 恭佑, 小池 崇文

概要 リアルタイムステレオキャプチャと両眼自動立体視ディスプレイの最小構成を用いた自動立体視3Dライブシステムを開発した。自動立体視3Dライブシステムと従来の2Dビデオ会議の表示内容を、定量分析と主観評価の両方を通じて比較した。定量的な評価として、アイトラッキングを用いて、会話相手の顔領域における固視時間を測定・評価した。本システムは、2Dビデオ会議と比較して、顔面領域での固視時間が長く、固視回数が多いことが示された。主観評価については、参加者を対象に調査を行い、その結果を分析した。ウィルコクソンの符号順位検定を用いて、2Dビデオ会議と我々のシステムを比較した。その結果、本システムは臨場感とリアリズムに対してより高い評価を達成した。

キーワード: 裸眼3Dディスプレイ, パララックスバリア, リアルタイムキャプチャ, アイトラッキング, 3Dビデオ会議

1. まえがき

テレプレゼンスとは、遠隔地にいる人々や風景が、あたかもシーンに存在するかのように、臨場感や没入感を与える技術である¹。2Dビデオ会議は、テレプレゼンスの一形態として、一般的に遠隔地にいる個人間のつながりを強化する。しかし、人間が3次元空間で知覚する場合、2次元ビデオ会議²を用いて真の臨場感を得ることは困難である。

田口らは、大型カメラアレイ⁴を用いたライトフィールドのリアルタイム取得と出力を実現したが、このようなセットアップには大きなハードウェア制約がある。プロジェクトStarline⁵とGoogle Beam⁶は、大型ライトフィールドディスプレイとマルチカメラキャプチャシステムによる高品質な対面コミュニケーションを可能にする。これらのシステムは忠実度の高い体験を提供する一方で、膨大なハードウェアと複雑な処理に依存しているため、実用的な展開には限界がある。さらに、対面および2Dビデオコミュニケーションに関する先行研究では、アイコンタクトの持続時間と頻度が対話者の印象にプラスの影響を与えることが示されている⁷。しかし、2Dディスプレイと自動立体視3Dディスプレイの顔指向性視線行動の比較は、まだ十分に検討されていない。

3Dテレプレゼンスをより実用的なものにするためのハードウェアオーバーヘッド¹⁰。先行する大規模システムやハードウェア集約型システムとは異なり、我々のアプローチは、リアルタイムの双方向3D通信に適した最小限のセットアップに焦点を当てている。さらに、ビデオ会議シナリオ用に設計されたビデオを用いて、顔に対する3つの視線メトリクスを具体的に評価した: (1)平均固視時間、(2)固視回数、(3)顔への最初の固視までの時間。

我々の知る限り、本研究は、最小限のハードウェア自動立体視3Dライブシステムを、顔指向の視線メトリクスと主観的評価を組み合わせた従来の2Dビデオ会議と定量的に比較した初期の研究の1つである。我々のアプローチには3つの貢献がある:

- ステレオカメラとアイトラッキングを用いた広視野自動立体視3Dディスプレイを用いた最小限の双方向リアルタイム通信システムを開発した。
- 我々は、2Dと自動立体視3Dディスプレイの両方で、静かなビデオ会議シナリオを参加者に提示し、3つの視線対顔のメトリクスを測定した: 平均固視時間、固視回数、最初の固視までの時間。
- 我々は、プレゼンスとリアリズムの主観的評価を実験し、自動立体視3Dライブシステムが2Dビデオ会議よりも大幅な改善をもたらすことを実証した。

視線接触とは無関係に自動立体

Received June 21, 2025; Revised October 10, 2025;
Accepted November 28

† Hosei University
(Tokyo, Japan)

†† RealImage Inc.
(Tokyo, Japan)

【機械翻訳コンテンツの著作権について】

当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
引用の際には、必ず原著論文の書誌情報をご記載ください。詳細は映像情報メディア学会著作権規定をご覧ください。

Development and Evaluation of Autostereoscopic 3D Live System with Real-Time Video Capture

Kyosuke Yanagida[†], Takafumi Koike (member)^{†,††}

Abstract We developed an autostereoscopic 3D live system using a minimal configuration of real-time stereo capture and a binocular autostereoscopic 3D display. We compared the displayed contents of the autostereoscopic 3D live system and conventional 2D video conferencing through both quantitative analysis and subjective evaluation. For the quantitative evaluation, we measured and assessed the fixation duration on the face area of the conversational partner using eye tracking. Our system showed longer fixation duration and a greater number of fixations on the face area compared to 2D video conferencing. For the subjective evaluation, we surveyed participants and analyzed the results. We used Wilcoxon signed-rank tests to compare our system with 2D video conferencing. The results show that our system achieves higher ratings for sense of presence and realism.

Key words: autostereoscopic 3D display, parallax barrier, real-time capture, eye-tracking, 3D video conferencing.

1. Introduction

Telepresence is a technology that gives a sense of presence and immersion to remote people and landscapes, as if they were present in the scene¹⁾. 2D video conferencing commonly strengthens connections between remote individuals as a form of telepresence. However, when humans perceive in three-dimensional space, it is challenging to achieve a true sense of presence using 2D video conferencing²⁾³⁾.

Taguchi *et al.* realized real-time acquisition and output of light fields using a large camera array⁴⁾, but such setups impose significant hardware constraints. Project Starline⁵⁾ and Google Beam⁶⁾ enable high-quality face-to-face communication through large light-field displays and multi-camera capture systems. While these systems provide high-fidelity experiences, their reliance on massive hardware and complex processing limits practical deployment. Moreover, previous studies on face-to-face and 2D video communication have shown that the duration and frequency of eye contact positively influence impressions of the interlocutor⁷⁾⁸⁾⁹⁾. However, comparisons of face-directed gaze behavior between 2D and autostereoscopic 3D displays remain insufficiently explored.

hardware overhead to make 3D telepresence more practical¹⁰⁾¹¹⁾. Unlike prior large-scale or hardware-intensive systems, our approach focuses on a minimal setup suitable for real-time, bidirectional 3D communication. In addition, we specifically evaluated three gaze metrics to face using a video designed for video conferencing scenarios: (1) average fixation duration, (2) number of fixations, and (3) duration until first fixation on the face.

To the best of our knowledge, this study is one of the early investigations that quantitatively compares a minimal hardware autostereoscopic 3D live system with conventional 2D video conferencing, combining face-directed gaze metrics with subjective evaluations. Our approach includes three contributions:

- We developed a minimal bidirectional real-time communication system using stereo cameras and a wide-view autostereoscopic 3D display with eye tracking.
- We presented participants with quiet video conferencing scenarios in both 2D and autostereoscopic 3D display, and measured three gaze-to-face metrics: average fixation duration, number of fixations, and time to first fixation.
- We experimented subjective evaluations of presence and realism, demonstrating that the autostereoscopic 3D live system yielded significant improvements over 2D video conferencing.

We experimented under conditions without mutual eye

Received June 21, 2025; Revised October 10, 2025; Accepted November 28, 2025

[†] Hosei University
(Tokyo, Japan)

^{††} ReallImage Inc.
(Tokyo, Japan)

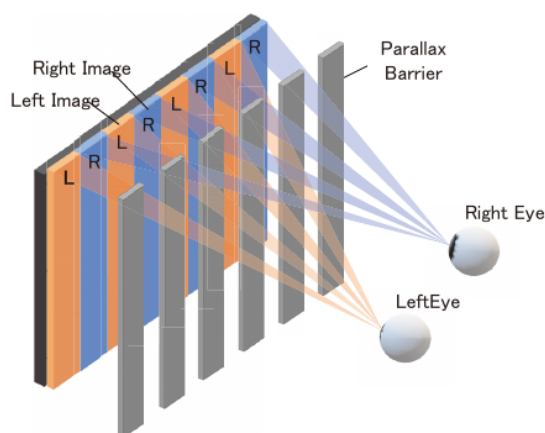


図1: 視差バリアシステム

視3Dライブシステムの効果を調べるため、相互視線接触のない条件下で実験を行った。

2. 関連研究

2.1 自動立体視3Dディスプレイ

自動立体視3Dディスプレイは、特殊なメガネやヘッドマウントディスプレイ(HMD)なしで立体視を行う。時間的多重法と空間的多重法に大別できる。時間多重法は、シャッターメガネと同期した左右の画像を交互に提示することで、高解像度の出力と柔軟な視聴位置を可能にするが、正確なタイミング制御が必要である¹²。一方、空間多重法では、視差バリアやレンチキュラーレンズ¹³⁾などのデバイスを用いて光学的に画像を分離する。視差バリアは、ピクセルを拡大するレンチキュラーレンズとは異なり、元のピクセル解像度を維持する利点がある。

視差バリアは、不透明で透明なスリットが交互に並ぶ垂直ストライプマスクである。図1は、2Dディスプレイに取り付けられた視差バリアが、左右の眼球画像を空間的に分割して立体視を実現する様子を示している。

視差バリアディスプレイの限界は、バリアがディスプレイに固定されているため、クロストークが発生しやすい狭い視野帯にある。これに対処するため、西村ら¹⁴は視点ベースの動的制御を提案し、中村ら¹⁵はドット空間のスリット中心で収束するサブピクセルを選択することでビューイングゾーンを拡張した。視点追跡と組み合わせた自動立体視表示方式を採用する。

2.2 3Dビデオシステムの評価

3Dディスプレイの評価には、クロストークや明るさの測定、主観的評価の収集、

2Dや3D視聴下での視線行動の分析などがしばしば含まれる。Liらは、視聴ゾーン全体のクロストークを評価する多次元的な方法を提案し、それが位置や視線によってどのように変化するかを示した¹⁷。山口らは、アイトラッキングでサブピクセルを検出し、その上に黒い画像を表示することで、視差バリアシステムの深度方向クロストークを抑制している¹⁸。これらの研究は、視点移動下での立体視品質を向上させるための重要な技術を提供する。

Lawrenceらは、自動立体視ディスプレイとマルチカメラキャプチャを組み合わせたリアルタイム3Dビデオ通信システム、Project StarlineとGoogle Beamを開発した⁵。Starlineは、2Dコールと比較して、プレゼンスとエンゲージメントの評価が有意に高いが、複雑な処理と制御された環境に依存しているため、一貫した品質が困難であると報告している。

他の研究では、3D画像における視線分布を分析している¹⁹。Häkkinenらは、2D映画の視聴者は主に人物に注目するのに対し、3Dでは複雑な構造や物体の近くに注目することを発見した²¹。しかし、これらの結果は、パートナーの顔が前景に提示されるビデオ会議の文脈では得られない。このように、対人コミュニケーションにおける視線行動は、まだ明らかにされていない。

2.3 コミュニケーションにおける視線とアイコンタクト

アイコンタクトは対人コミュニケーションにおいて重要な役割を果たす。パートナーの顔面の固視時間が長いと、信頼が高まり、よりスムーズな情報交換が容易になることが示されている⁷。顔に直接注視することで、相手の魅力が増すことも報告されている⁸。さらに、実験的研究により、ヒトは非常に短い視線提示からでも「見る」ことを迅速に知覚できることが実証されている⁹。この知見は、インタラクションの最初に瞬間的なアイコンタクトが、興味や親和性の感情を引き起こし、それがコミュニケーションの全体的な経過に影響を与える可能性があることを示唆している。

これらの知見は、遠隔コミュニケーションにおいても同様に重要である²²。しかし、2Dビデオ会議やテレプレゼンスシステムでは、ディスプレイとカメラの間の変位が相互注視の実現を困難にしている。これまでの研究で、ハーフミラーを使ってキャプチャカメラの光軸をディスプレイの中心に仮想的に合わせることで、2Dビデオ会議でのアイコンタクトが親密感を高めることができることが示されている²³。

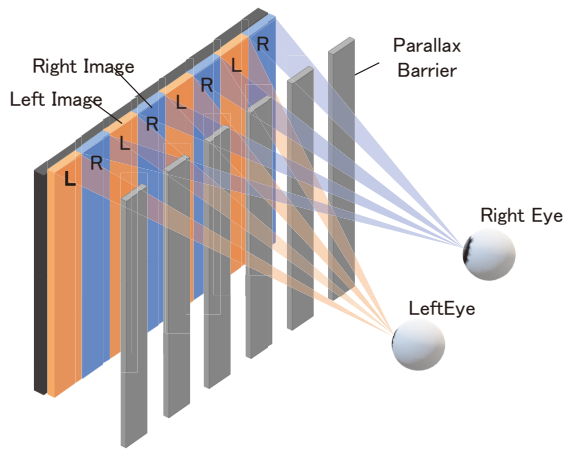


Fig. 1: Parallax barrier system.

contact to examine the effects of the autostereoscopic 3D live system independently of eye contact.

2. Related Work

2.1 Autostereoscopic 3D Displays

Autostereoscopic 3D displays provide stereoscopic viewing without special glasses or a head-mounted display (HMD). They can be broadly classified into temporal and spatial multiplexing methods. The temporal multiplexing method alternately presents left and right images synchronized with shutter glasses, enabling high-resolution output and flexible viewing positions, but requiring precise timing control¹²⁾. In contrast, the spatial multiplexing method separates images optically using devices such as parallax barriers or lenticular lenses¹³⁾. Parallax barriers have the advantage of maintaining the original pixel resolution, unlike lenticular lenses that magnify pixels.

A parallax barrier is a vertically striped mask with alternating opaque and transparent slits. Fig. 1 shows how a parallax barrier attached to a 2D display spatially divides left- and right-eye images to achieve stereoscopic viewing.

A limitation of parallax barrier displays is their narrow viewing zone, where crosstalk easily occurs because the barrier is fixed to the display. To address this, Nishimura *et al.*¹⁴⁾ proposed viewpoint-based dynamic control, and Nakamura *et al.*¹⁵⁾ expanded the viewing zone by selecting sub-pixels that converge at slit centers in the dot-space. We adopt the autostereoscopic display method combined with viewpoint tracking.

2.2 Evaluation of 3D Video Systems

Evaluation of 3D displays often involves measuring crosstalk and brightness, collecting subjective eval-

uations, or analyzing gaze behavior under 2D and 3D viewing. Li *et al.* propose a multidimensional method to evaluate crosstalk across the viewing zone and showed how it varies with position and gaze¹⁷⁾. Yamaguchi *et al.* suppress depth-direction crosstalk in a parallax barrier system by detecting subpixels with eye tracking and displaying black images on them¹⁸⁾. These studies provide key techniques for improving stereoscopic quality under viewpoint shifts.

Lawrence *et al.* developed Project Starline and Google Beam, real-time 3D video communication systems combining autostereoscopic displays with multi-camera capture⁵⁾⁶⁾. They report significantly higher presence and engagement ratings for Starline compared to 2D calls, though its reliance on complex processing and controlled environments makes consistent quality difficult.

Other studies analyze gaze distribution in 3D images¹⁹⁾²⁰⁾. Häkkinen *et al.* found that viewers of 2D movies focused mainly on people, whereas in 3D they looked more at complex structures and near objects²¹⁾. However, these results are not obtained in a video conferencing context, where a partner's face is presented in the foreground. Thus, gaze behavior in interpersonal communication settings remains to be clarified.

2.3 Gaze and Eye Contact in Communication

Eye contact is a crucial role in interpersonal communication. Longer fixation durations on a partner's face have been shown to enhance trust and facilitate smoother information exchange⁷⁾. Directly gazing at the face has also been reported to increase the perceived attractiveness of the partner⁸⁾. Furthermore, experimental studies have demonstrated that humans can rapidly perceive "being looked at" even from very brief gaze presentations⁹⁾. This finding suggests that a momentary eye contact at the beginning of an interaction can trigger feelings of interest and affinity, which may subsequently influence the overall course of communication.

These findings are equally important in remote communication²²⁾. However, in 2D video conferencing and telepresence systems, the displacement between the display and the camera makes mutual gaze difficult to achieve. Previous studies have shown that, by using a half mirror to virtually align the optical axis of the capture camera with the center of the display, eye contact in 2D video conferencing can enhance a sense of intimacy²³⁾. We quantitatively examine the effects of

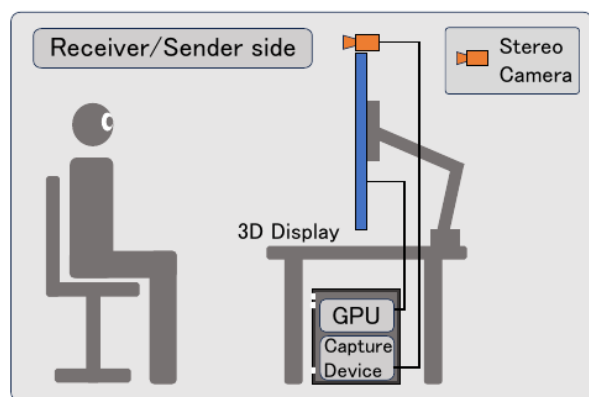


図2: 自動立体視による3Dライブシステム構成。

視線行動に対する2D提示と自動立体視による3D提示の効果を定量的に検討した。具体的には、静かな環境で録画されたビデオ会議映像を使用し、参加者に2Dビデオ会議と自動立体視3Dライブシステムの両方を通してそれを視聴させ、顔領域における3つの視線指標(平均固視時間、固視回数、顔への最初の固視までの時間)を測定させた。

3. 自動立体視3Dライブシステム

3.1 概要

自動立体視3Dライブシステムのハードウェア構成とソフトウェア処理の概要を説明する。図2に自動立体視3Dライブシステムの構成を示す。自動立体視3Dディスプレイ、ステレオカメラ、PCの3つのデバイスから構成される。自動立体視3Dディスプレイは、中村ら¹⁵が開発した視点追従を用いた視差バリアシステムを採用している。SDIケーブルがステレオカメラをキャプチャデバイスに接続する。

本システムは、ステレオ映像キャプチャ、ステレオ映像通信、自動立体視3Dディスプレイ上のステレオ映像ディスプレイの順に処理を行う。ステレオカメラは非圧縮映像を出力し、それをキャプチャする。システムはキャプチャした映像をGPUに転送し、UDP通信プロトコルを用いてステレオ映像を送信する。システムは受信したステレオ映像を、観察者の視点移動に対応する左右の目の画像とともに表示した。この構成は最も単純であると考えられる。

3.2 ステレオカメラの配置

本システムでは、キーストーンの歪みを最小化し、会議シナリオに対して安定した再現性を提供するため、記録に並列キャプチャ法を採用した。

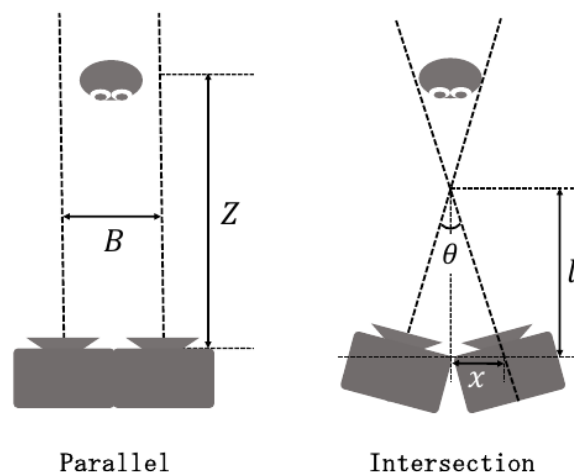


図3: ステレオ画像キャプチャのための並列・交差法の概念図。

自動立体視3Dディスプレイは、立体視を実現するために、両眼視差のあるステレオ画像を入力画像として必要とする。ステレオ画像を撮影する方法として、カメラの光軸を平行に揃える平行法と、左右のカメラの光軸を収束させる交差法の2つが存在する。図3は、並列法と交差法の概念図である。平行法 d_p の視差は、式(1)を用いて計算した。ここで、 l は左右のカメラレンズの中心間の距離、 f はカメラの焦点距離、 Z は被写体までの距離を表す。

$$d = \frac{l \cdot f}{Z}. \quad (1)$$

並列法は、被写体までの距離に反比例した視差を作成する。ビデオ会議シナリオにおいて、観察者がカメラに近づくと、視差が大きくなり、立体視が難しくなる。

交差法により、自動立体視3Dディスプレイは、カメラの光軸がディスプレイ平面で交差する距離 l に位置する被験者を再現することができる。オブジェクトは光交差点の前方では自動立体視3D表示面から飛び出し、光交差点では自動立体視3D表示面の後方にオブジェクトが現れる。カメラの収束角度を制御することで、飛び出す量を制御することが可能である。 x はステレオカメラのベースラインを表し、 θ は表示平面自体で被写体を再現する設定 $l = Z$ を表す。ここで、 x はステレオカメラのベース

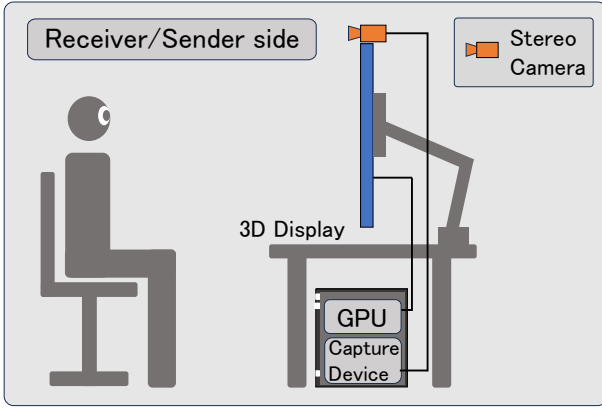


Fig. 2: Autostereoscopic 3D live system configuration.

2D versus autostereoscopic 3D presentation on gaze behavior. Specifically, we used video conferencing footage recorded in a quiet environment and had participants view it through both 2D video conferencing and the autostereoscopic 3D live system, measuring three gaze metrics on the facial region: average fixation duration, number of fixations, and time to first fixation on the face.

3. Autostereoscopic 3D Live System

3.1 Overview

We provide an overview of the hardware configuration and software processing of the autostereoscopic 3D live system. Fig. 2 shows the configuration of the autostereoscopic 3D live system. It consists of three devices: the autostereoscopic 3D display, the stereo camera, and the PC. The autostereoscopic 3D display adopts the parallax barrier system using viewpoint tracking developed by Nakamura *et al.*¹⁵⁾. SDI cables connect the stereo camera to a capture device.

Our system performs the following processing in order: stereo video capture, stereo video communication, and stereo video display on the autostereoscopic 3D display. The stereo camera outputs uncompressed video, which is then captured. The system transfers the captured video to the GPU and transmits stereo video using the UDP communication protocol. The system displayed the received stereo video with the left and right eye images corresponding to the observer's viewpoint movement. We consider this configuration to be the simplest.

3.2 Stereo Camera Arrangement

We adopt a parallel capture method for recording in our system, because it minimizes keystone distortion and provides stable reproduction for conferencing scenarios. The autostereoscopic 3D display requires

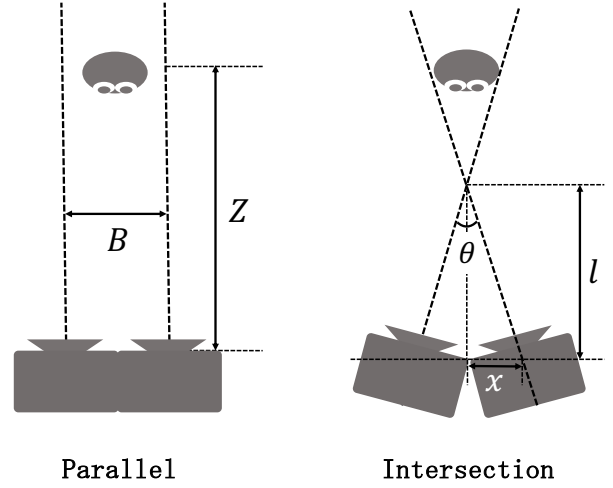


Fig. 3: Conceptual diagram of parallel and intersection methods for stereo image capture.

stereo images with binocular disparity as input images to achieve stereoscopic viewing. Two methods exist for capturing stereo images: the parallel method, which aligns the camera optical axes in parallel, and an intersection method, which converges the left and right camera optical axes. Fig. 3 shows a conceptual diagram of the parallel and intersection methods. We calculated the parallax for the parallel method d_p using equation (1), where I represents the distance between the centers of the left and right camera lenses, f denotes the camera's focal length, and Z indicates the distance to the subject.

$$d = \frac{I \cdot f}{Z}. \quad (1)$$

The parallel method creates a parallax inversely proportional to the distance to the subject. When the observer moves closer to the camera in video conferencing scenarios, the parallax increases, making stereoscopic viewing more challenging.

The intersection method allows the autostereoscopic 3D display to reproduce subjects positioned at the distance l , where the optical axes of the cameras intersect at the display plane. The objects pop out of the autostereoscopic 3D display plane in front of the optical intersection point, whereas the objects appear behind the autostereoscopic 3D display plane in the optical intersection point. It is possible for us to control the amount of popping out by controlling the camera convergence angle. We calculated the distance l from the starting points of the left and right camera optical axes to their intersection point using equation (2), where x represents the stereo camera baseline and θ denotes the

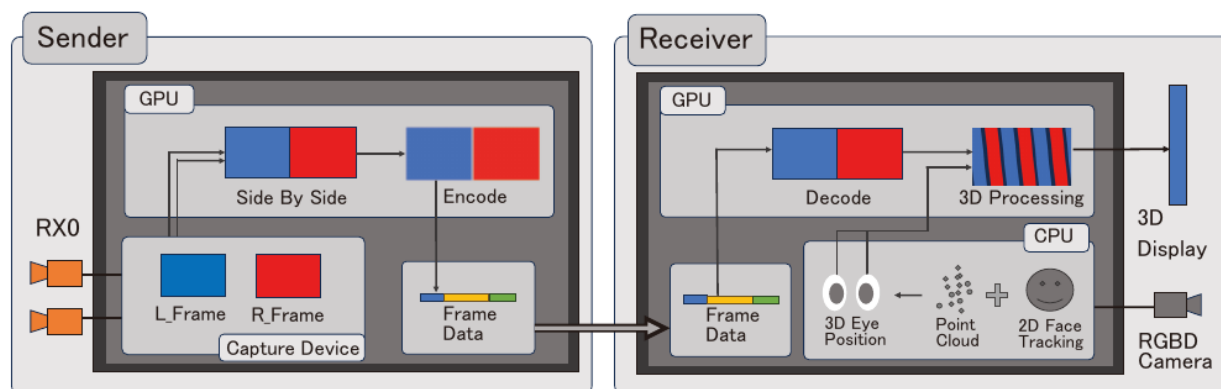


図4: 送信側と受信側の画像処理と通信の流れ。

カメラの光軸間の収束角を測定した。

$$l = x \tan \frac{\theta}{2}. \quad (2)$$

ラインを表し、 θ は表示平面自体で被写体を再現することを表す。

立体視が困難な観察者であっても、突出量を制御することで立体視を容易にした。ビデオ会議のための光軸を設定し、被写体をディスプレイ平面の後ろに配置する $l < Z$ を選択することで、リアリズムをさらに向上させる。しかし、収束角が大きいとキーストーンの歪みが生じるため、現実的な再現のために歪みの少ない並列捕捉法を採用した。

ステレオカメラを使用する場合、視線を合わせることが難しい。そこで、図2に示すプロトタイプを用いて、視線計測による主観評価を行った。被験者にカメラへのフォーカスを指示した動画を使用した。この問題は、将来、光学系を使ってステレオカメラレンズとディスプレイ中心を仮想的に位置合わせすることで解決できるだろう。

3.3 実装方法

RealImage社のRealImage 3D-32TM (3D-32)を使用。視差バリアシステムを用いた自動立体視3Dディスプレイで、解像度は3840×2160ピクセル。最大リフレッシュレートは60Hzである。3D-32は、モアレパターンとクロストークを最小限に抑えながらイメージを提供するように設計されている。RealImage Inc.は、ディスプレイの上にトラッキングカメラを固定し、カラー画像から観察者の顔情報を取得し、深度画像から対応する深度情報を取得する。

つのSONY DSC-RX0(RX0)カメラを利用する。RX0は1080p60までのHDMI出力をサポートしている。

寸法は幅59.0mm、高さ40.5mm、深さ29.8mmである。RX0はレンズを前面の中心に配置するため、2台のRX0カメラを並べて配置すると、ベースラインは約60mmとなる。人間の平均的な瞳孔間距離は約60mmであるため、2台のRX0カメラを使用することで、両眼視差の再現が可能になる。

PCはIntel(R) CoreTM i9-12900KのCPUと64GBのメモリを搭載している。GPUとしてNVIDIA RTX A4000を使用。Unity(バージョン2021.3.21)を用いて、自動立体視3Dディスプレイの制御を実装する。ビデオキャプチャには、Unity上で実行可能なフレームワークであるAVPro DeckLinkを使用する。そこで、Blackmagic Design DeckLink Duo 2をキャプチャデバイスとして使用する。通信には、UnityRenderStreamingを使用し、WebRTCをUnity内で処理できるように実装している。

図4に本システムにおける通信処理の概要を示す。PCは、送信側をエンコードしながら、リアルタイムでステレオ映像をサイドバイサイドでキャプチャする。送信PCにインストールされたDeckLink Duo 2キャプチャカードは、RX0からのビデオをキャプチャする。送信PCは左右の眼球画像を1つのサイドバイサイドフレームにまとめ、受信PCに送信する。受信側PCは送信側からフレームデータを受信し、デコードする。また、RGB-Dカメラは点群とRGBフレームを撮影する。受信PCは、RGBフレームと点群を用いて、2次元顔位置から3次元眼球位置を計算する。最後に、受信側PCは受信側に自動立体視3Dディスプレイを表示した。自動立体視3Dディスプレイでは、受信側で取得した観察者の3D視点位置とともにフレームデータを使用する。図5は送信側で撮影したステレオ映像、図6は受信側の自動立体視3Dディスプレイに表示したステレオ映像である。

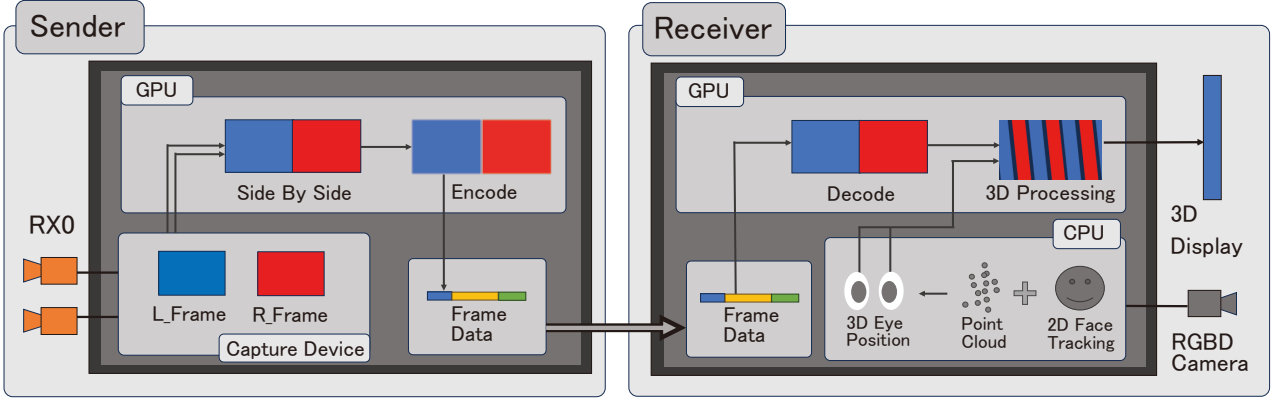


Fig. 4: Image processing and communication flow on the sending and receiving sides.

convergence angle between the camera optical axes.

$$l = x \tan \frac{\theta}{2}. \quad (2)$$

Setting $l = Z$ reproduces the subject on the display plane itself.

We facilitated stereoscopic viewing by controlling the amount of protrusion, even for observers with difficulty in stereoscopic perception. Realism is further enhanced by setting the optical axes for video conferencing and choosing $l < Z$, which placed the subject behind the display plane. However, a large convergence angle caused keystone distortion, so we adopted the parallel capture method with less distortion for realistic reproduction.

It is difficult to align the line of sight when using a stereo camera. Therefore, we used the prototype shown in Fig. 2 for the subjective evaluation with gaze measurements. We used videos where the subject was instructed to focus on the camera. This issue could be resolved in future work by using an optical system to virtually align the stereo camera lenses with the display center.

3.3 Implementation

We use RealImage 3D-32TM (3D-32) from RealImage Inc. It is the autostereoscopic 3D display using the parallax barrier system, with a resolution of 3840×2160 pixels. It has a maximum refresh rate of 60Hz. The 3D-32 is designed to provide imagery while minimizing moire patterns and crosstalk. RealImage Inc. fixes a tracking camera above the display, where it acquires the observer's facial information from the color image and the corresponding depth information from the depth image.

We utilize two SONY DSC-RX0 (RX0) cameras. The RX0 supports HDMI output up to 1080p60. The dimensions are 59.0 mm in width, 40.5 mm in height, and

29.8 mm in depth. The RX0 places its lens at the center of the front face, so placing two RX0 cameras side by side creates a baseline of approximately 60mm. As the average human interpupillary distance is approximately 60 mm, using two RX0 cameras enables the reproduction of binocular disparity.

The PC has Intel(R) CoreTM i9-12900K CPU and 64GB of memory. We used NVIDIA RTX A4000 as GPU. We implement the control of the autostereoscopic 3D display using Unity (version 2021.3.21). For video capture, we use AVProDeckLink, a framework that can run on Unity. Therefore, we use Blackmagic Design DeckLink Duo 2 as the capture device. For communication, we use UnityRenderStreaming, which implements WebRTC to enable handling it within Unity.

Fig. 4 shows an overview of the communication processing in our system. PC captures stereo video in real-time, side-by-side image generation, encoding on the sending side. The DeckLink Duo 2 capture card installed in the sending PC captures the video from the RX0s. The sending PC combines the left and right eye images into a single side-by-side frame and sends it to the receiving PC. The receiving PC receives frame data from the sending side and decodes it. Also, the RGB-D camera captures point clouds and an RGB frame. The receiving PC calculates 3D eye positions from a 2D face position using the RGB frame and the point clouds. Finally, the receiving PC displayed the autostereoscopic 3D display on the receiving side. We use the frame data along with the observer's 3D viewpoint position acquired on the receiving side for the autostereoscopic 3D display. Fig. 5 shows the stereo video captured on the sending side, and Fig. 6 shows the stereo video displayed on the autostereoscopic 3D display on the receiving side.



図5:送信側で撮影されたステレオ画像。



図6:受信側の自動立体視3Dで表示された立体視映像の両目の位置からキャプチャされたステレオ画像。

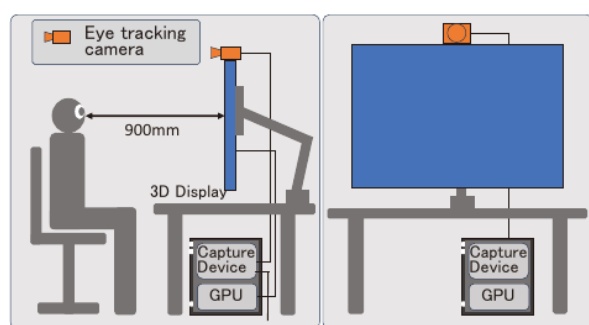


図7:顔面の固視時間を測定するためのシステム構成 (左:側面図、右:正面図)。

4. 顔領域の視線行動の比較

4.1 実験システム

アイコンタクトと、誰かの顔を見るという行為は、非常に重要な役割を果たした。しかし、多くの2Dビデオ会議システムはアイコンタクトの認識を困難にし、通信を制限した²⁴。

我々は、自動立体視3D映像が人間の人物に対する視線行動にどのような影響を与えるかを調べるために、3つの定量的指標を実験した。最初の指標は平均固視時間であり、これは各固視がスクリーンに表示された顔面上でどのくらい続いたかを示す。2つ目の指標は、ビデオのオンセットから視聴者の顔への最初の固視までの時間として測定される、最初の固視までの時間であった。3つ目の指標は、視線が顔の特定の部分に戻る頻度を反映する固視回数である。これらの指標は、顔に対する視線の発生と注意の関与の程度の両方を総合的に測定したものである。

ユーザエンゲージメントと感情反応

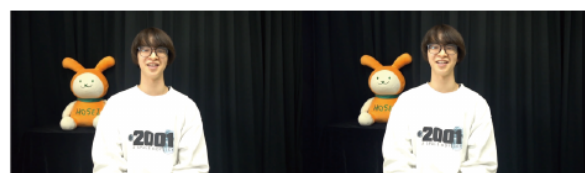


図8:実験に使用したステレオビデオの一部。

を評価するためにユーザビリティ研究で一般的に使用されている²⁶、これらのメトリクスは、3Dビデオが対人コミュニケーションをいかに効果的にサポートするかを評価するためにも不可欠である。そこで、自動立体視3Dコミュニケーションにおけるアイコンタクトの役割を検討するために、2D条件と3D条件の固視時間と固視回数を比較した。

図7は、自動立体視3Dディスプレイ、視線追跡カメラ、PCを含む、我々のシステム構成を示している。自動立体視3Dディスプレイは、視線追跡カメラが上部に設置されていた。自動立体視3Dディスプレイ上に2Dと3Dのコンテンツを表示した。そして、実験システムを用いて各映像を観察するために、被験者を最適視距離(OVD)付近に座らせた。3Dビデオコンテンツから左目のみを用いて2Dコンテンツを生成した。自動立体視3Dディスプレイではパラレルメソッドビデオを使用し、2Dディスプレイではパラレルメソッドビデオの左目ビデオを使用した。図8は、被験者を中心とした2つのRX0を使用して、1分間のビデオ撮影を行い、ビデオ会議をシミュレートしたものである。

4.2 視線解析手法

我々は、Baltrušaitisらによって開発されたオープンソースの機械学習ベースの視線追跡システムであるOpenFaceを使用した¹⁶。OpenFaceはビデオベースの視線追跡システムであり、視線追跡カメラで撮影されたビデオを入力することで、カメラ座標系における観察者の視点座標と視線ベクトルを得ることができる。

カメラ座標系における左目の視点の3次元座標 x_L, y_L, z_L と視線ベクトル成分 i_L, j_L, k_L を用いて、左目の視線とディスプレイ平面の交点を表す表示座標 W_L と H_L を式(3)、(4)のように推定した。右目の視線ベクトルとディスプレイ平面の交点も同様に決定することができる。

$$W_L = x + \frac{z}{k_L} \cdot i_L. \quad (3)$$

$$H_L = y + \frac{z}{k_L} \cdot j_L. \quad (4)$$

固視点の決定方法については、研究者の間で継続的な議論が



Fig. 5: Stereo Images captured on the sending side.



Fig. 6: Stereo Images captured from the position of both eyes of a stereoscopic video displayed in autostereoscopic 3D on the receiving side.

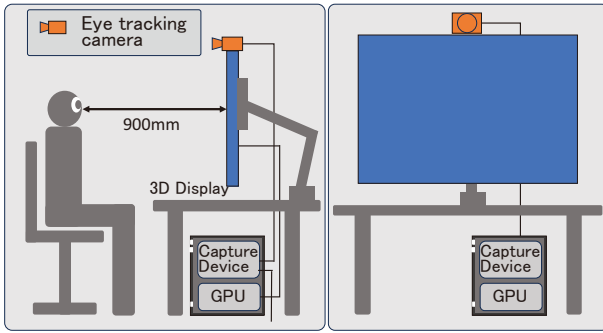


Fig. 7: System configuration for measuring fixation duration on the face (left: side view, right: front view).

4. Comparison of Face Area Gazing Behavior

4.1 Experiment System

Eye contact and the act of looking at someone's face played extremely important roles. However, many 2D video conferencing systems made recognizing eye contact difficult, which restricted communication²⁴⁾.

We experimented with three quantitative metrics to investigate how autostereoscopic 3D video influences gaze behavior toward human figures. The first metric was average fixation duration, which indicates how long each fixation lasted on the face displayed on the screen. The second metric was duration until first fixation, measured as the time from the onset of the video to the viewer's first fixation on the face. The third metric was the number of fixations, which reflects how frequently the gaze returned to a specific part of the face. These metrics collectively measured both the occurrence of gaze toward the face and the extent of attentional engagement.

Commonly used in usability studies to assess user en-



Fig. 8: Part of the stereo video used in the experiment.

gagement and emotional response²⁶⁾, these metrics are also essential for evaluating how effectively 3D video supports interpersonal communication. Therefore, we compared fixation duration and fixation count between 2D and 3D conditions to examine the role of eye contact in an autostereoscopic 3D communication.

Fig. 7 shows our system configuration, which includes an autostereoscopic 3D display, a gaze tracking camera, and a PC. The autostereoscopic 3D display had the gaze tracking camera at the top. We displayed 2D and 3D content on the autostereoscopic 3D display. And we seated the subjects near the optimal viewing distance (OVD) to observe each video using the experimental system. We generated the 2D content using only the left-eye view from the 3D video content. For the autostereoscopic 3D display, we used the parallel method video, while for the 2D display, we used the left-eye video from the parallel method video. Fig. 8 shows that a one-minute video was captured using two RX0s centered on the subject, simulating a video conference.

4.2 Gaze Analysis Method

We used OpenFace, an open-source, machine learning-based gaze tracking system developed by Baltrušaitis *et al.*^{16) 25)}. OpenFace is a video-based gaze tracking system that, by inputting video captured with a gaze tracking camera, can obtain the observer's viewpoint coordinates and gaze vectors in the camera coordinate system.

We estimated the display coordinates W_L and H_L , which represent the intersection points of the left eye's gaze and the display plane, using the 3D coordinates x_L, y_L, z_L of the left eye's viewpoint in the camera coordinate system and the gaze vector components $\vec{i}_L, \vec{j}_L, \vec{k}_L$, as shown in equations (3) and (4). The intersection of the right eye's gaze vector and the display plane could be similarly determined.

$$W_L = x + \frac{z}{k_L} \cdot \vec{i}_L. \quad (3)$$

$$H_L = y + \frac{z}{k_L} \cdot \vec{j}_L. \quad (4)$$

There has been an ongoing discussion among researchers regarding the method for determining the fix-

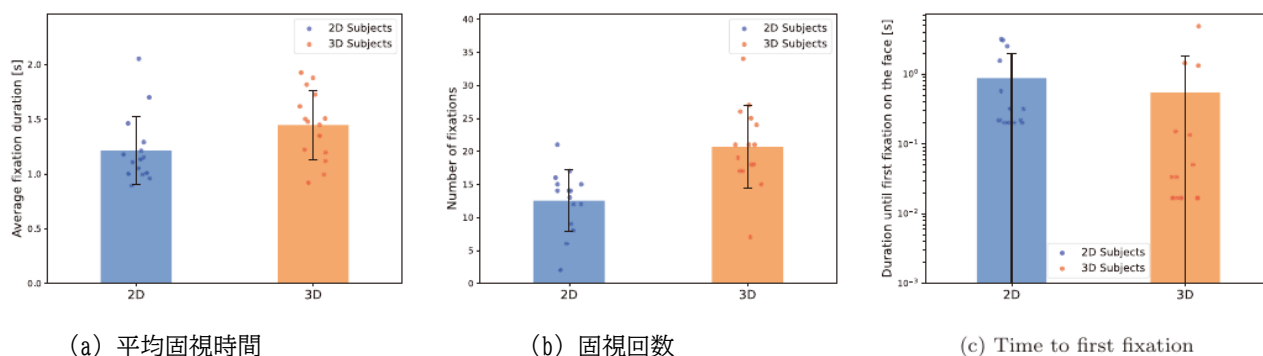


図9:2Dおよび3Dディスプレイにおける顔領域の視線メトリクスの結果(各参加者について示す)。棒グラフは各条件の平均値を表し、エラーバーは標準偏差を表す。

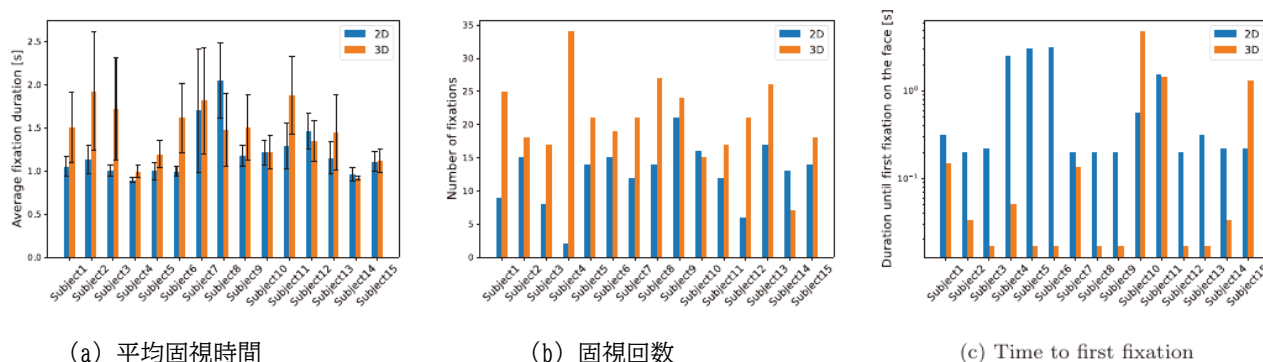


図10:2Dおよび3Dディスプレイにおける顔領域の視線メトリクスの結果(各参加者について示す)。平均固視時間のエラーバーは標準偏差を示す。

表1:2Dビデオ会議と自動立体視3Dライブシステムにおける視線メトリクスのWilcoxon符号順位検定による評価結果。

	statistic	p-value
平均固視時間	19.0	0.0090
フィクセーション数	8.0	0.0007
初回固定までの時間	26.0	0.0554

なされている。長松らは、視点位置と視線ベクトルが既知で、視線ベクトルがディスプレイなどの物体と交差する場合、物体の水平成分の midpoint を固定点として較正することで、誤差が最小になることを示した²⁷。固視点(W, H)を、両目の位置から発生する視線ベクトルとディスプレイの交点の midpoint と定義する。

まず、各ビデオフレーム内の顔領域を検出し、顔の固視時間を測定した。Googleが開発したライブラリMediaPipe²⁸を用いて、顔の位置とそのランドマークを推定し、顔の対応する特定部分を抽出した。

注視点(W, H)が検出された顔のバウンディングボックス内に留まる累積時間を固視時間と定義した。注視点領域外に移動するまで、システムは連続的に固視時間を蓄積した。このような連続的な蓄積は、それぞれ顔への1回の固視としてカウントした。

4.3 結果

表1は、3つの視線メトリクスに対するウィルコクソンの符号順位検定の結果をまとめたものである。平均固視時間と固視回数については、2Dディスプレイと自動立体視3Dディスプレイの間に有意差が認められた。図9は、各表示条件における視線メトリクスの平均値と標準偏差を示したものである。また、図10は、両条件下での全参加者の視線メトリクスの個別結果を示している。図10において、平均固視時間のエラーバーは標準偏差を示す。

平均固視時間については、参加者は一般に、自動立体視3Dディスプレイの下で、より長い固視を示した。2D条件ではより長い持続時間を示す個体もいたが、3D条件では全体的な分布がより長い固視に有意にシフトした(統計=19.0、 $p=0.0090$)。このことは、自動立体視による3Dディスプレイが、

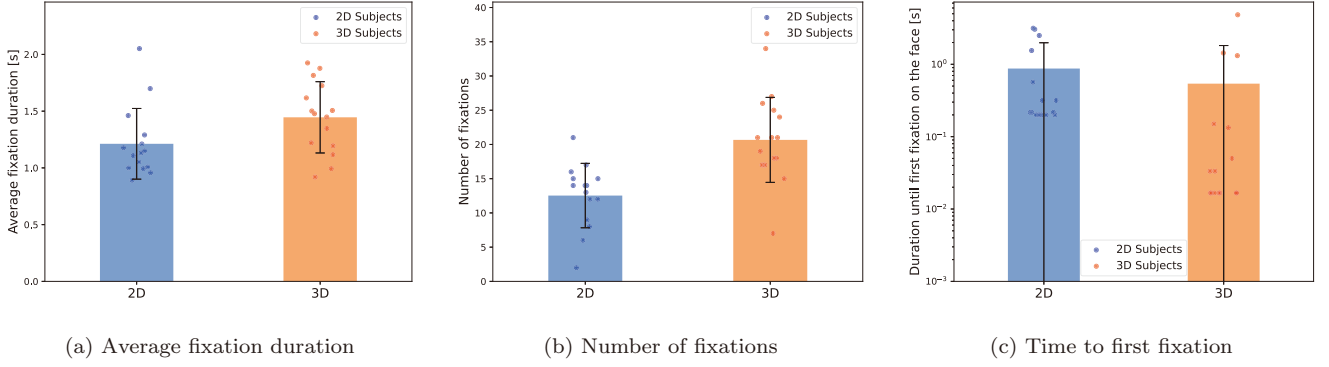


Fig. 9: Results of gaze metrics for the facial region on 2D and 3D displays, shown for each participant. Bars represent the mean values for each condition, and error bars indicate standard deviation.

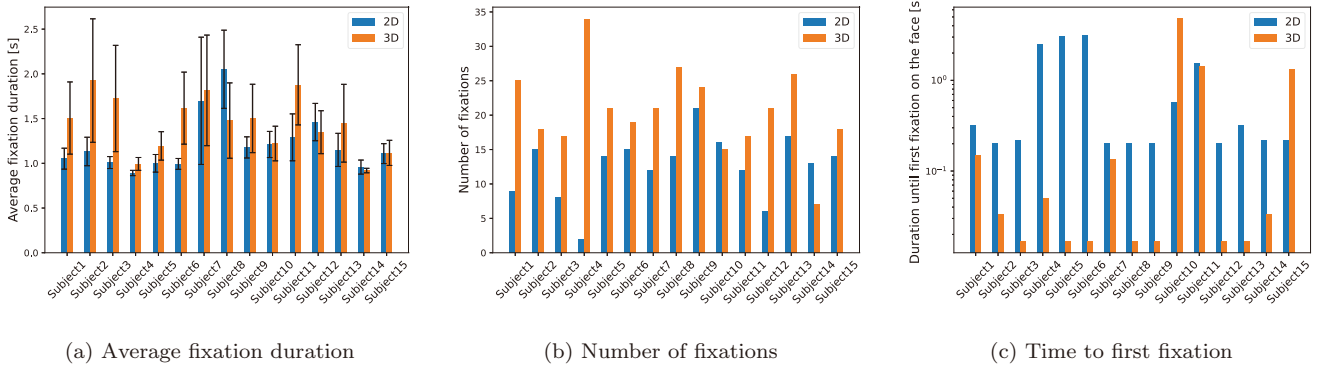


Fig. 10: Results of gaze metrics for the facial region on 2D and 3D displays, shown for each participant. Error bars in the average fixation duration indicate standard deviation.

Table 1: Evaluation results using Wilcoxon signed-rank tests for gaze metrics in 2D video conferencing and autostereoscopic 3D live systems.

	statistic	<i>p</i> -value
Average fixation duration	19.0	0.0090
Number of fixations	8.0	0.0007
Time-to-first-fixation	26.0	0.0554

ation point. Nagamatsu *et al.* demonstrated that if the viewpoint position and gaze vector are known and the gaze vector intersects with an object such as a display, calibrating the midpoint of the horizontal component of the object as the fixation point results in minimal error²⁷⁾. We define the fixation point (W, H) as the midpoint of the intersection points of the gaze vectors, originating from the positions of both eyes, with the display.

We first detected the face area within each video frame to measure fixation duration on the face. We used MediaPipe²⁸⁾, a library developed by Google, to estimate the position of the face and its landmarks and to extract the corresponding specific part of the face. We defined fixation duration as the cumulative time

during which the gaze point (W, H) stayed within the bounding box of the detected face. The system accumulated fixation duration continuously until the gaze point moved outside the region. We counted each such continuous accumulation as one fixation on the face.

4.3 Result

Table 1 summarizes the results of the Wilcoxon signed-rank tests for the three gaze metrics. Significant differences were found between the 2D and autostereoscopic 3D displays for average fixation duration and number of fixations. Fig. 9 shows the means and standard deviations of the gaze metrics for each display condition. And, Fig. 10 shows the individual results of the gaze metrics for all participants under both conditions. In Fig. 10, the error bars for average fixation duration indicate the standard deviation.

For average fixation duration, participants generally exhibited longer fixations under the autostereoscopic 3D display. Although some individuals showed longer durations in the 2D condition, the overall distribution was significantly shifted toward longer fixations in the 3D condition (statistic = 19.0, $p = 0.0090$). This indicates that the autostereoscopic 3D display promoted

パートナーの顔へのより安定した持続的な注意を促進したことを示している。

固視回数については、自動立体視3Dディスプレイは2Dディスプレイと比較して有意に高いカウント数を示した(統計=8.0, $p=0.0007$)。このことは、参加者がより長い固視を維持するだけでなく、より頻繁に顔に戻ることを示唆しており、立体視下での顔の手がかりの顕著性がより強く反映されている。

一方、顔面への初回固視までの時間は、条件間で有意差は認められなかった(統計=26.0, $p=0.0554$)。このように、参加者が最初に視線を顔に向ける速度は、立体視提示の影響をあまり受けなかった。その代わりに、最初の視線固定の後に違いが現れ、その後の視線の持続時間と頻度に影響を与えた。

4.4 視線行動に関する考察

Häkkinenらは、2Dビデオと比較した3Dビデオでは、視聴者は複雑な構造や人物の前にある物体に視線を集中させる傾向があると報告している²¹。しかし、我々の実験で使用したビデオは、ビデオ会議のような前景に大きく座っている個人を特徴としており、これが視線集中の一因となっている可能性が高い。さらに、我々はビデオ会議のような単純な環境での視線追跡に焦点を当てたが、奥行きのある環境では、視線は奥行きを伝えるオブジェクトに集中し、自動立体視3Dディスプレイの特徴である奥行き知覚を活用すると予想した。

さらに、個々の固視時間のばらつきや、ビデオベースのトラッキングに内在する確率的なノイズを潜在的な外乱として考慮する必要がある。したがって、本実験では、2Dディスプレイと自動立体視3Dディスプレイの顔面固視時間には有意な差があったが、その差を単一の要因に帰することは決定的にできなかった。したがって、我々の実験の主な発見は、ビデオ会議のような静かな環境では、3Dビデオは2Dビデオと比較して、被験者の顔に対してより多くの視線を引き寄せる傾向があるということであった。

2Dディスプレイと自動立体視3Dディスプレイの間で、顔面への最初の固定までの時間に統計的に有意な差は見られなかった。この結果から、ディスプレイの形式は、ビデオ開始後の顔に対する初期の視覚的方向性にほとんど影響を与えないことが示唆された。外的な注意散漫をなくすため、制御された静かな環境で実験を行った。

また、3D画像における奥行きの知覚は、人物の存在だけでなく、周囲の物体や背景要素にも影響されることが認識された。顔の知覚の効果を分離するために、簡略化した1つのシーンを使用した。

5. 主観的評価

5.1 アンケート評価

我々は、2Dビデオ会議システムと我々のシステムの両方で実験し、参加者が報告した主観的評価を検証した。参加者は、各システムを使って5分間のミーティングを2回に分けて行った。参加者が両方のシステムを使用する順番をランダムにした。従来の2Dビデオ会議システムは、ステレオカメラ入力を単眼カメラ入力に置き換えるという修正を利用した。

アイコンタクトは、プレゼンスと相互理解を高めることで、対面コミュニケーションにおいて重要な役割を果たす。しかし、我々のプロトタイプシステムは、ステレオカメラがユーザーの光軸に沿うのではなく、ディスプレイの上に配置されているため、真のアイコンタクトを達成することができない。主観的評価を解釈する際には、この限界を考慮する必要がある。本実験では、自動立体視3Dディスプレイのみによって提供される臨場感と臨場感を検証するために、アイコンタクトのない条件下での臨場感と臨場感を評価することを目的とする。

参加者は、各会議後にリッカート尺度を用いて3つの指標を評価した。各手法について、臨場感、リアルさ、会話の焦点を評価した。プレゼンスとは、遠隔地のコミュニケーション相手と同じ場所にいること的感覺である。リアリズムとは、遠隔地にいるコミュニケーション相手、つまりディスプレイに映る相手と対面しているという感覺のことである。集中度は、各条件における会話の焦点である。評価は、1(「全くない」)から5(「非常にある」)までの5段階のリッカート尺度を用いて評価した。さらに、アンケートでは、参加者に、2つの方法間のコミュニケーションにおいて認識した差異を記述した自由記述回答を求めた。

5.2 結果

参加者は20代の15人で、男性10人、女性5人であった。このうち、7人が眼鏡を、5人がコンタクトレンズを、3人が矯正されていない視力であった。さらに、3名の参加者が非点収差を有していた。

各評価指標について、Wilcoxon 符号順位検定を用いて分析した。

more stable and sustained attention to the partner's face.

For the number of fixations, the autostereoscopic 3D display yielded significantly higher counts compared to the 2D display (statistic = 8.0, $p = 0.0007$). This suggests that participants not only maintained longer fixations, but also returned to the face more frequently, reflecting stronger salience of facial cues under stereoscopic viewing.

In contrast, the time to first fixation on the face showed no significant difference between conditions (statistic = 26.0, $p = 0.0554$). Thus, the speed with which participants initially directed their gaze toward the face was not substantially affected by stereoscopic presentation. Instead, differences emerged after the first fixation, influencing the duration and frequency of subsequent gazes.

4.4 Discussion on Gaze Behavior

Häkkinen *et al.* reported that, in 3D videos compared to 2D videos, viewers tended to focus their gaze on objects located in front of complex structures or people²¹). However, the videos used in our experiment featured individuals seated prominently in the foreground, such as in a video conference setting, which likely contributed to focused gaze. Additionally, while we focused our experiment on gaze tracking in a simple environment similar to video conferences, in environments with depth, we expected gaze to concentrate on objects that convey depth, leveraging the depth perception characteristic of the autostereoscopic 3D displays.

Furthermore, we should consider variations in individual fixation duration and probabilistic noise inherent in video-based tracking as potential disturbances. Therefore, although there was a significant difference in fixation duration on the face between 2D and the autostereoscopic 3D displays in our experiment, we couldn't definitively attribute the difference to any single factor. Hence, the main finding of our experiment was that in a quiet environment similar to video conferences, 3D video tended to attract more gaze towards the subject's face compared to 2D video.

We found no statistically significant difference in the duration until first fixation on the face between the 2D and the autostereoscopic 3D displays. The result suggests that the display format had minimal influence on participants initial visual orientation toward the face after the video began. We experimented in a controlled, quiet environment to eliminate external distractions. We also recognized that the perception of depth in 3D

images can be affected not only by the presence of people but also by surrounding objects and background elements. We used a single, simplified scene to isolate the effect of facial perception.

5. Subjective Evaluation

5.1 Questionnaire Evaluation

We experimented with both a 2D video conferencing system and our system to examine the subjective evaluations reported by participants. Participants took part in two separate 5-minute meetings using each system. We randomized the order in which participants used both systems. The conventional 2D video conferencing system utilized a modification where the stereo camera input of our system was replaced with a monocular camera input.

Eye contact is a crucial role in face-to-face communication by enhancing presence and mutual understanding. However, our prototype system cannot achieve true eye contact, as the stereo cameras are positioned above the display instead of being aligned with the user's optical axis. This limitation should be considered when interpreting the subjective evaluation. In this experiment, we aim to evaluate presence and realism under conditions without eye contact, in order to examine the sense of presence and realism provided solely by the autostereoscopic 3D display.

Participants evaluated three metrics using the Likert scale after each meeting. They rated the sense of presence, realism, and the conversational focus for each method. The sense of presence is the feeling of being in the same place as the remote communication partner. Realism is the feeling of being face-to-face with the remote communication partner, i.e., the person who appears on the display. The concentration is conversational focus for each condition. We evaluated the ratings using a 5-point Likert scale ranging from 1 ("Not at all") to 5 ("Very much"). Additionally, the questionnaire asked participants to provide free-text responses describing any differences they perceived in communication between the two methods.

5.2 Result

The participants consisted of 15 individuals in their twenties, including 10 males and 5 females. Among them, seven participants wore glasses, five wore contact lenses, and three had uncorrected vision. Additionally, three participants had astigmatism.

We analyzed each evaluation metric using a Wilcoxon signed-rank test. The null hypothesis stated that

表2:2Dビデオ会議と自動立体視3Dライブシステムのリッカート尺度による評価結果。

	Mean(Standard deviation)		
	2D videoconferencing	our system	p-value
プレゼンス	2.33(1.11)	3.53(.92)	< 0.01
Realism	2.00(.85)	3.93(1.10)	< 0.001
濃度	3.73(.80)	4.07(1.22)	> 0.05

帰無仮説は、“我々の自動立体視3Dライブシステムと従来の2Dビデオ会議との間で、評価指標の分布に有意差はない”と述べている。我々は、対立仮説を“我々の自動立体視3Dライブシステムと従来の2Dビデオ会議との間で、評価基準の分布に有意な差がある”と述べている。

表 2 に調査結果を示す。我々は、2Dビデオ会議と自動立体視3Dライブシステムとの間に、臨場感とリアリズムの点で有意な違いがあることを発見した。しかし、濃度に有意な差は見られなかった。

数名の参加者が、自由記述の回答で同様の意見を表明した。主に2つの肯定的なコメントがあった：まず、2人の参加者が、“パンチングやカメラの方を向いているようなジェスチャーをすることで、臨場感が高まる”と述べた。第二に、3人の参加者が、“自動立体視3Dライブシステムを使用することで、被験者の奥行き知覚が可能になり、ボディランゲージがより魅力的になる”と指摘した。否定的な面では、3人の参加者が“ディスプレイを見てアイコンタクトを取ろうとするときでさえ、実際のアイコンタクトは達成されなかった”と述べた。これらの回答は、参加者のシステム経験に共通するテーマを浮き彫りにした。

5.3 主観的評価に関する考察

この結果から、臨場感やリアルさに対する評価が高いのは、パラレル撮影法で撮影された画像の立体視に起因すると考えられ、これにより、視聴者は自動立体視3Dディスプレイにおける会話相手の深さを知覚することができる。

なお、今回の結果は、相互注視がない場合に得られたものである。相互注視は社会的存在感とコミュニケーションの質をさらに高めることが知られているため、我々の知見は、自動立体視3Dライブシステムの潜在的な利点の下限推定値である。さらに、臨場感の平均評価がリアリズムに比べて低いのは、フラットディスプレイの使用に起因している。

洞窟自動仮想環境(CAVE)やHMDのような没入型技術は、視聴者が3D空間内に存在すると感じることを可能にし、臨場感を高めることが報告されている。会話の焦点に有意な差がないことから、視覚的な3Dプレゼンテーションは会話への関与に最小限の影響しか与えない可能性が示唆された。

参加者の中には、2Dビデオ会議システムでも我々のシステムでもアイコンタクトがとれないと言及する人もいた。私たちのシステムは、キャプチャシステムをディスプレイの上に配置し、それによって直接のアイコンタクトを防いだ。実験結果は、自動立体視3Dライブシステムは、2Dビデオ会議よりも対面でのインタラクションに近いコミュニケーション体験を提供し、その結果、全体的なエンゲージメントが高くなることを示した。

6. むすび

我々は、ステレオキャプチャと自動立体視3Dディスプレイの最も単純な構成で、自動立体視3Dライブシステムを開発した。その有効性を評価するために、定量的・主観的な評価を用いて、従来の2Dビデオ会議と比較した。ビデオ会議シナリオのために特別に設計されたビデオを用いて、視線行動を測定・分析した。さらに、参加者は主観評価中に両方のシステムを経験した。その結果、自動立体視3Dライブシステムは、2Dビデオ会議と比較して、臨場感や臨場感について高い評価を受けていることが示された。

いくつかの課題が残っている。まず、自動立体視3Dディスプレイの上にユーザー向けカメラが取り付けられているため、アイコンタクトが難しい。さらに、適切な奥行き知覚を保証するために、ベースラインの長さや収束角度に基づく視聴者固有のキャリブレーションが必要である。

References

- 1) M. Minsky : “Telepresence,” Omni, pp. 45–51 (1980)
- 2) M. Lombard and T. B. Ditton : “At the Heart of It All: The Concept of Presence,” Journal of Computer-Mediated Communication, 3, 2 (1997)
- 3) D. P. Hove and B. Watson : “The Shortcomings of Video Conferencing Technology, Methods for Revealing Them, and Emerging XR Solutions,” Presence: Virtual and Augmented Reality, 31, 3, pp.283–305 (2022)
- 4) Y. Taguchi, T. Koike, K. Takahashi, and T. Naemura : “TransCAIP : A Live 3D TV System Using a Camera Array and an Integral Photography Display with Interactive Control of Viewing Parameters,” IEEE Transactions on Visualization and Computer Graphics, 15, 5, pp. 841–852 (2009)
- 5) J. Lawrence, D. B. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong : “Project Starline : A High-Fidelity Telepresence System,” ACM Transactions on Graphics (Proc. SIGGRAPH Asia), 40 (2021)

Table 2: Evaluation results using Likert scale for 2D video conferencing and autostereoscopic 3D live systems.

	Mean(Standard deviation)		
	2D videoconferencing	our system	p-value
Sense of presence	2.33(1.11)	3.53(.92)	< 0.01
Realism	2.00(.85)	3.93(1.10)	< 0.001
Concentration	3.73(.80)	4.07(1.22)	> 0.05

“There is no significant difference in the distribution of the evaluation metrics between our autostereoscopic 3D live system and conventional 2D video conferencing.” We stated the alternative hypothesis as “There is a significant difference in the distribution of evaluation criterion between our autostereoscopic 3D live system and conventional 2D video conferencing.”

Table 2 shows the survey results. We found significant differences between 2D video conferencing and the autostereoscopic 3D live system in terms of sense of presence and realism. However, there was no significant difference in concentration.

Several participants expressed similar opinions in their free-text responses. There were mainly two positive comments: First, two participants mentioned that “Making gestures such as punching or pointing towards the camera enhances the sense of presence.” Second, three participants noted that “Using the autostereoscopic 3D live system allows a sense of depth perception for the subject, making body language more engaging.” On the negative side, three participants mentioned that “Even when attempting to make eye contact by looking at the display, actual eye contact was not achieved.” These responses highlighted common themes in participants experiences with the system.

5.3 Discussion on Subjective Evaluation

Based on the results, we considered that the high ratings for sense of presence and realism were attributable to the stereoscopic viewing of images captured using the parallel filming method, which enabled viewers to perceive the depth of their conversational partner in the autostereoscopic 3D display.

It should be noted that the present results were obtained in the absence of mutual gaze. Since mutual gaze is known to further increase social presence and communication quality, our findings represent a lower-bound estimate of the potential benefit of autostereoscopic 3D live system. Additionally, the lower average rating for sense of presence compared to realism was attributed to the use of a flat display. Immersion technologies such

as the Cave automatic virtual environment (CAVE) or HMD, which allow viewers to feel present within a 3D space, have been reported to enhance the sense of presence. The lack of significant difference in conversational focus suggested that visual 3D presentation may have minimal impact on conversational engagement.

Some participants mentioned that they could not make eye contact in both the 2D video conferencing system and our system. Our system positioned its capturing system above the display, thereby preventing direct eye contact. The experimental results indicated that the autostereoscopic 3D live system offered a communication experience more similar to face-to-face interaction than 2D video conferencing, resulting in higher overall engagement.

6. Conclusion

We developed the autostereoscopic 3D live system with the simplest configuration of stereo capture and the autostereoscopic 3D display. To evaluate its effectiveness, we compared it with conventional 2D video conferencing using quantitative and subjective assessments. We measured and analyzed gaze behavior using videos specifically designed for video conferencing scenarios. Furthermore, participants experienced both systems during the subjective evaluation. The results indicated that the autostereoscopic 3D live system received higher ratings for sense of presence and realism compared to the 2D video conferencing.

We have several remaining challenges. First, making eye contact is difficult because the user-facing camera is mounted above the autostereoscopic 3D display. Additionally, viewer-specific calibration based on baseline length and convergence angle is needed to ensure proper depth perception.

References

- 1) M. Minsky : “Telepresence,” *Omni*, pp. 45–51 (1980)
- 2) M. Lombard and T. B. Ditton : “At the Heart of It All: The Concept of Presence,” *Journal of Computer-Mediated Communication*, 3, 2 (1997)
- 3) D. P. Hove and B. Watson : “The Shortcomings of Video Conferencing Technology, Methods for Revealing Them, and Emerging XR Solutions,” *Presence: Virtual and Augmented Reality*, 31, 3, pp.283–305 (2022)
- 4) Y. Taguchi, T. Koike, K. Takahashi, and T. Naemura : “TransCAIP : A Live 3D TV System Using a Camera Array and an Integral Photography Display with Interactive Control of Viewing Parameters,” *IEEE Transactions on Visualization and Computer Graphics*, 15, 5, pp. 841–852 (2009)
- 5) J. Lawrence, D. B. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong : “Project Starline : A High-Fidelity Telepresence System,” *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40 (2021)

- 6) Google : “Google Beam,” <https://starline.google> (Accessed: June 19, 2025)
- 7) H. Kreysa, L. Kessler and S. R. Schweinberger : “Direct Speaker Gaze Promotes Trust in Truth-Ambiguous Statements,” *PLOS ONE*, 11(9), e0162291 (2016)
- 8) R. E. Kaisler and H. Leder : “Trusting the Looks of Others: Gaze Effects of Faces in Social Settings,” *Perception*, 45(8), pp. 875–892 (2016)
- 9) Y. Nakajima, Y. Tani, Y. Isogaya and T. Sato : “The Influence of Duration on the Perception of Gaze Direction and Being Looked at,” *Japanese Journal of Psychonomic Science*, 29, pp. 69–70 (2010)
- 10) K. Yanagida, T. Koike, G. Hamagishi, and H. Takahashi : “Real-Time 3D Video Communication System Using Stereo Video Capture and Autostereoscopic Display,” *Proc. of International Display Workshops, 3DSAp1-6* (2023)
- 11) K. Yanagida, T. Koike, G. Hamagishi, and H. Takahashi : “Real-Time 3D Video Communication System Using Stereo Video Capture and Autostereoscopic Display with Eye Tracking,” *Proc. of IEEE International Conference on Image Electronics and Visual Computing* (2024)
- 12) L. F. Hodges : “Tutorial: Time-Multiplexed Stereoscopic Computer Graphics,” *IEEE Comput. Graph. Appl.*, 12, 2, pp. 20–30 (1992)
- 13) T. Okoshi : “Three-Dimensional Imaging Techniques,” Atara Press (2011)
- 14) T. Nishimura, G. Hamagishi, K. Kusafuka, K. Yoshimoto, and H. Takahashi : “Autostereoscopic Display with Wide Viewing Zone Using Binocular Image Control Method in Divided Area Based on Viewing Position,” *Optical Engineering*, 60, 12 (2021)
- 15) H. Nakamura, G. Hamagishi, K. Yoshimoto, H. Takahashi, T. Matsumoto, and K. Kusafuka : “A Novel Eye Tracking System to Expand Viewing Area in all Directions for Glasses-Free 3D Display Displayable in both Portrait and Landscape Modes,” *Proc. of International Display Workshops* (2018)
- 16) T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency : “OpenFace 2.0: Facial Behavior Analysis Toolkit,” *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 59–66 (2018)
- 17) S. Li and L. Cao : “Multidimensional Crosstalk Analysis in Autostereoscopic Displays: Integrating Subjective and Objective Evaluations for Image Quality Assessment,” *Optics Express*, 33, 8, pp. 16911–16928 (2025)
- 18) Y. Yamaguchi, K. Yamamoto, H. Yamaguchi, Y. Yamaguchi, H. Nakamura, G. Hamagishi, K. Yoshimoto, T. Matsumoto, K. Kusafuka, and H. Takahashi : “Further Crosstalk Reduction Method with Eye-Tracking for Glasses-Free Stereoscopic Display in Both Portrait and Landscape Modes,” *Proc. of International Display Workshops, 3D6/3DSA6-2* (2019)
- 19) C. J. Lin, R. Widyaningrum, and Y. T. Prasetyo : “The Effect of 3D TVs on Eye Movement and Motor Performance,” *Applied Sciences*, 13, 4, p. 2656 (2023)
- 20) S. Wibirama and K. Hamamoto : “3D Gaze Tracking on Stereoscopic Display Using Optimized Geometric Method,” *Transactions on Electronics, Information and Systems*, 134, 3 (2014)
- 21) J. Häkkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman : “What Do People Look at When They Watch Stereoscopic Movies?,” *Proc. of SPIE - The International Society for Optical Engineering*, 7524 (2010)
- 22) J. O. Hietanen, M. J. Peltola and J. K. Hietanen : “Psychophysiological Responses to Eye Contact in a Live Interaction and in Video Call,” *Psychophysiology*, 57(6), e13587 (2020)
- 23) N. Kaiser, K. Henry and H. Eyjólfssdóttir : “Eye Contact in Video Communication: Experiences of Co-creating Relationships,” *Frontiers in Psychology*, 13, 852692 (2022)
- 24) M. Chen : “Leveraging the Asymmetric Sensitivity of Eye Contact for Videoconferencing,” *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*, 8, pp. 49–56 (2002)
- 25) E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling : “Rendering of Eyes for Eye-Shape Registration and Gaze Estimation,” *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 3756–3764 (2015)
- 26) T. Tullis and B. Albert : “Behavioral and Physiological Metrics,” *Measuring the User Experience (2nd Edition)*, Morgan Kaufmann, pp. 163–186 (2013)
- 27) T. Nagamatsu, J. Kamahara, and N. Tanaka : “Calibration-Free Gaze Tracking Using a Binocular 3D Eye Model,” *CHI '09 Ex-*

tended Abstracts on Human Factors in Computing Systems, 6, pp. 3613–3618 (2009)

- 28) V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann : “BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs,” *CoRR*, abs/1907.05047 (2019)



Kyosuke Yanagida received the B.S. in computer and Information science from Seikei University in 2022 and the M.S. in computer and information sciences from Hosei University in 2024. He is currently a doctoral course student at Graduate School of Computer and Information Sciences, Hosei University.



Takafumi Koike received his Ph.D. in information and communication engineering from The University of Tokyo. He was formerly a senior researcher at Yokohama Research Laboratory, Hitachi, Ltd. Since 2013, he has been a Professor at the Faculty of Computer and Information Sciences at Hosei University. He is also a founder and CEO of RealImage Inc. His research interests include media technologies connected between the real and cyber worlds, such as light field displays/cameras, computer graphics, and virtual/augmented/mixed reality.

- 6) Google : “Google Beam,” <https://starline.google> (Accessed: June 19, 2025)
- 7) H. Kreysa, L. Kessler and S. R. Schweinberger : “Direct Speaker Gaze Promotes Trust in Truth-Ambiguous Statements,” *PLOS ONE*, 11(9), e0162291 (2016)
- 8) R. E. Kaisler and H. Leder : “Trusting the Looks of Others: Gaze Effects of Faces in Social Settings,” *Perception*, 45(8), pp. 875–892 (2016)
- 9) Y. Nakajima, Y. Tani, Y. Isogaya and T. Sato : “The Influence of Duration on the Perception of Gaze Direction and Being Looked at,” *Japanese Journal of Psychonomic Science*, 29, pp. 69–70 (2010)
- 10) K. Yanagida, T. Koike, G. Hamagishi, and H. Takahashi : “Real-Time 3D Video Communication System Using Stereo Video Capture and Autostereoscopic Display,” *Proc. of International Display Workshops, 3DSAp1-6* (2023)
- 11) K. Yanagida, T. Koike, G. Hamagishi, and H. Takahashi : “Real-Time 3D Video Communication System Using Stereo Video Capture and Autostereoscopic Display with Eye Tracking,” *Proc. of IEEE International Conference on Image Electronics and Visual Computing* (2024)
- 12) L. F. Hodges : “Tutorial: Time-Multiplexed Stereoscopic Computer Graphics,” *IEEE Comput. Graph. Appl.*, 12, 2, pp. 20–30 (1992)
- 13) T. Okoshi : “Three-Dimensional Imaging Techniques,” Atara Press (2011)
- 14) T. Nishimura, G. Hamagishi, K. Kusafuka, K. Yoshimoto, and H. Takahashi : “Autostereoscopic Display with Wide Viewing Zone Using Binocular Image Control Method in Divided Area Based on Viewing Position,” *Optical Engineering*, 60, 12 (2021)
- 15) H. Nakamura, G. Hamagishi, K. Yoshimoto, H. Takahashi, T. Matsumoto, and K. Kusafuka : “A Novel Eye Tracking System to Expand Viewing Area in all Directions for Glasses-Free 3D Display Displayable in both Portrait and Landscape Modes,” *Proc. of International Display Workshops* (2018)
- 16) T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency : “OpenFace 2.0: Facial Behavior Analysis Toolkit,” *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 59–66 (2018)
- 17) S. Li and L. Cao : “Multidimensional Crosstalk Analysis in Autostereoscopic Displays: Integrating Subjective and Objective Evaluations for Image Quality Assessment,” *Optics Express*, 33, 8, pp. 16911–16928 (2025)
- 18) Y. Yamaguchi, K. Yamamoto, H. Yamaguchi, Y. Yamaguchi, H. Nakamura, G. Hamagishi, K. Yoshimoto, T. Matsumoto, K. Kusafuka, and H. Takahashi : “Further Crosstalk Reduction Method with Eye-Tracking for Glasses-Free Stereoscopic Display in Both Portrait and Landscape Modes,” *Proc. of International Display Workshops, 3D6/3DSA6-2* (2019)
- 19) C. J. Lin, R. Widyaningrum, and Y. T. Prasetyo : “The Effect of 3D TVs on Eye Movement and Motor Performance,” *Applied Sciences*, 13, 4, p. 2656 (2023)
- 20) S. Wibirama and K. Hamamoto : “3D Gaze Tracking on Stereoscopic Display Using Optimized Geometric Method,” *Transactions on Electronics, Information and Systems*, 134, 3 (2014)
- 21) J. Häkkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman : “What Do People Look at When They Watch Stereoscopic Movies?,” *Proc. of SPIE - The International Society for Optical Engineering*, 7524 (SPIE)
- 22) J. O. Hietanen, M. J. Peltola and J. K. Hietanen : “Psychophysiological Responses to Eye Contact in a Live Interaction and in Video Call,” *Psychophysiology*, 57(6), e13587 (2020)
- 23) N. Kaiser, K. Henry and H. Eyjólfssdóttir : “Eye Contact in Video Communication: Experiences of Co-creating Relationships,” *Frontiers in Psychology*, 13, 852692 (2022)
- 24) M. Chen : “Leveraging the Asymmetric Sensitivity of Eye Contact for Videoconferencing,” *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*, 8, pp. 49–56 (2002)
- 25) E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling : “Rendering of Eyes for Eye-Shape Registration and Gaze Estimation,” *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 3756–3764 (2015)
- 26) T. Tullis and B. Albert : “Behavioral and Physiological Metrics,” *Measuring the User Experience (2nd Edition)*, Morgan Kaufmann, pp. 163–186 (2013)
- 27) T. Nagamatsu, J. Kamahara, and N. Tanaka : “Calibration-Free Gaze Tracking Using a Binocular 3D Eye Model,” *CHI '09 Ex-*

tended Abstracts on Human Factors in Computing Systems, 6, pp. 3613–3618 (2009)

- 28) V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann : “BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs,” *CoRR*, abs/1907.05047 (2019)



Kyosuke Yanagida received the B.S. in computer and Information science from Seikei University in 2022 and the M.S. in computer and information sciences from Hosei University in 2024. He is currently a doctoral course student at Graduate School of Computer and Information Sciences, Hosei University.



Takafumi Koike received his Ph.D. in information and communication engineering from The University of Tokyo. He was formerly a senior researcher at Yokohama Research Laboratory, Hitachi, Ltd. Since 2013, he has been a Professor at the Faculty of Computer and Information Sciences at Hosei University. He is also a founder and CEO of RealImage Inc. His research interests include media technologies connected between the real and cyber worlds, such as light field displays/cameras, computer graphics, and virtual/augmented/mixed reality.