Available online at www.sciencedirect.com
SCIENCE DIRECT®
J. Vis. Commun. Image R. 17 (2006) 490–508
JOURNAL OF VISUAL Communication & IMAGE Representation
www.elsevier.com/locate/jvci

Stitching of H.264 video streams for continuous presence multipoint videoconferencing

Ashish K. Banerji, Kannan Panchapakesan, Kumar Swaminathan *

Hughes Network Systems, 11717 Exploration Lane, Germantown, MD 20876, USA

Received 30 April 2004; accepted 4 May 2005
Available online 18 August 2005

Abstract

This paper proposes a video stitching method applicable to continuous presence multipoint videoconferencing. The proposed approach is universally applicable yet computationally simpler than a pixel-domain approach which is computationally expensive due to the need for full re-encoding. A compressed-domain approach is far simpler but its applicability is limited to H.261. The proposed method employs a blend of compressed-domain and pixel-domain tools to eliminate the drift that results when a compressed-domain approach is attempted for H.263 and H.264. Coding parameters from incoming streams are reused on the outgoing side, thereby avoiding the computational bottleneck of re-estimating them. Essential details for implementing the proposed method are presented for H.264, along with simulation results for validating it. Application of the drift-free stitching method to H.263 is provided as an Appendix A.

Keywords: Videoconferencing; Continuous presence multipoint; Video stitching; Video combining; H.264; H.263; Compressed-domain; Pixel-domain; Drift-free; Intra prediction; Inter prediction

* Corresponding author.
E-mail address: kswami@hns.com (K. Swaminathan).

1047-3203/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2005.05.007

1. Introduction

In the past, videoconferencing has been an expensive and uncommon service. However, with the availability of low-priced components, significant development in video coding and networking technologies, and an ever increasing demand for remote collaboration, its popularity is on the rise. Videoconferencing is emerging as a promising tool for corporate meetings, distance learning, sales, telecommuting, telemedicine, remote testimony in legal environments, and several other applications. Traditionally, videoconferencing services have been deployed in a point-to-point scenario, where the audio–video information from one end-point is presented to the opposite end-point. This is a straightforward proposition since the video monitor at each end-point needs to only display the single image from the other end-point.

A more useful scenario is multipoint videoconferencing, where multiple video images from multiple end-points must somehow be displayed on a single video monitor so that a participant at one location can simultaneously see and hear the participants at diverse locations. A multipoint videoconferencing service typically includes a multipoint control unit (MCU), which receives the audio–video streams from each end-point, processes these, and distributes a suitable audio–video stream to the other end-points. The audio processing typically involves decoding the audio streams, summing them, and re-encoding the sum as one audio signal. However, video processing is much more difficult since there is no direct way to sum the input video streams. Hence, a multipoint videoconferencing session is commonly configured in one of two operating modes—the first is known as *voice activated* (VA) or *switched presence* mode, wherein one video source is selected based on the current audio level and is sent to all the terminals except itself (the video source being broadcast receives the video from the previous speaker). The second is *continuous presence* (CP) mode, wherein the MCU uses a *video stitcher* (also called *video combiner*) that is responsible for stitching the input video streams from multiple end-points to generate a *single* video stream, which when decoded by another end-point produces images from multiple remote end-points spatially composed and displayed on the video monitor of the local end-point. For example, if there are five or fewer end-points in a video conference, the four (or fewer) remote end-points may be displayed simultaneously in a 2×2 array, with each occupying one quadrant of the combined picture. In the case where there are more than five end-points, either a display partitioning different from 2×2 can be used; or else, one of the four quadrants of the 2×2 partitioning, such as the lower right quadrant, may be configured for VA operation along with the other three static quadrants.

There are two traditional approaches to performing video stitching—the *pixel-domain approach* and the *compressed-domain approach*. In the pixel-domain approach, the incoming streams are fully decoded to pixel-domain, and a spatially composed version of these is re-encoded. Although this can be implemented irrespective of the coding standard used, the re-encoding step is computationally complex and memory intensive, and adds to the end-to-end delay in a video conference. High latency severely affects interactive user experience and so end-to-end delay is a