*Proc. SPIE Vol. 2411 Human Vision, Visual Processing, and Digital Display VI,* (San Jose, CA), pp. 120–127, Feb. 1995.

# On video and audio data integration for conferencing

Thrasyvoulos N. Pappas

Signal Processing Research Department
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

Raynard O. Hinds[1]

Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139

## ABSTRACT

In video conferencing applications, the perceived quality of the video signal is affected by the presence of an audio signal (speech). To achieve high compression rates, video coders must compromise image quality in terms of spatial resolution, gray-scale resolution, and frame rate and may introduce various kinds of artifacts. We consider tradeoffs in gray-scale resolution and frame rate, and use subjective evaluations to assess the perceived quality of the video signal in the presence of speech. In particular, we explore the importance of lip synchronization.

In our experiment we used an original gray-scale sequence at QCIF resolution, 30 frames/second, and 256 gray levels. We compared the 256-level sequence at different frame rates (1 to 7.5 frames/sec) with a two-level version of the sequence at 30 frames/sec. The viewing distance was 20 image heights, or roughly two feet from an SGI workstation. We used uncoded speech. To obtain the two-level sequence we used an adaptive clustering algorithm for segmentation of video sequences. The binary sketches it creates move smoothly and preserve the main characteristics of the face, so that it is easily recognizable. More importantly, the rendering of lip and eye movements is very accurate. The test results indicate that, when the frame rate of the full gray-scale sequence is low (less than 5 frames/sec), most observers prefer the two-level sequence.

## 1. INTRODUCTION

The perceived quality of a video signal can be affected by the presence of an audio signal. In order to achieve high compression rates, video coders must compromise image quality in terms of spatial resolution, frame rate, and gray-scale (or color) resolution and may also introduce various kinds of coding artifacts (spatial artifacts, motion artifacts, or blurring due to motion). The goal of this paper is to examine how some of these compromises affect the perceived image quality in the presence of an audio signal.

We consider video conferencing and video phone applications, where the video data is a head and shoulder sequence and the audio data is speech. We consider tradeoffs in gray-scale resolution and temporal resolution (frame rate), and use subjective evaluations to assess the perceived image quality in the presence of speech. In particular, we explore the importance of the synchronization of lip movements with speech. The tradeoffs we consider in this paper are intended to provide guidelines for the design of a video compression scheme as well as for the choice of a display device.

In our experiments, we compared video sequences at high gray-scale resolution (256 gray levels), low spatial resolution (QCIF), and low frame rates (1 to 7.5 frames/sec) with a sequence at high temporal resolution (30 frames/sec), the same low spatial resolution (QCIF), and *minimal* gray-scale resolution (two gray levels). All sequences were obtained from the same original sequence. We used uncoded speech of relatively high quality. The viewing distance was 20 image heights, or roughly two feet from an SGI workstation.

[1] Consultant at AT&T Bell Laboratories, Summer 1994.

The low frame rate sequences were obtained by decimating the original sequence in time. This was done to preserve the spatial resolution of each frame. The two-level sequence was obtained using an adaptive clustering algorithm for segmentation of video sequences [1]. It is an extension of an adaptive clustering algorithm for segmentation of still images that uses spatial constraints and takes into consideration the local intensity characteristics of the image [2]. The three-dimensional algorithm uses temporal constraints and temporal local intensity adaptation to ensure that motion is smooth. It produces a binary (two-level) sketch of the image sequence which preserves the movements of the head, the eyes, and the lips, while other details in the background and clothing may be lost. The moving sketches also preserve the main characteristics of the face, so that it is easily recognizable.

The test results indicate that, when the frame rate of the full gray-scale sequence is low (less than 5 frames/sec), most observers prefer the binary sequence that preserves smooth motion and lip synchronization. About 25% of the observers prefer the “clear” (full gray-scale) images to the binary sketches at all rates, indicating that the tradeoff we chose may have been a little too extreme. This actually strengthens the results because it indicates that the two-level sequence is quite objectionable. Therefore, an observer would choose it only if the jerkiness and lack of lip synchronization of the grey-scale sequence at a low frame rate is really annoying. Thus, our test indicates that 5 frames/sec is probably a critical rate for video conferencing. Lower rate sequences are very objectionable and people are ready to make significant compromises to avoid them.

The remainder of this paper is organized as follows. In Section 2, we review the adaptive clustering algorithm for segmentation of video sequences. In Section 3, we present the subjective evaluation test. The conclusions are summarized in Section 4.

## 2. VIDEO SEQUENCE SEGMENTATION

In this section, we review the adaptive clustering algorithm for obtaining a binary sketch of a sequence of gray-scale images [1]. It is an extension of the adaptive clustering algorithm for still images [2]. It uses a Bayesian approach to segment the gray-scale images into black and white regions. Each segmented image preserves significant features while discarding unimportant detail. It is an adaptive thresholding scheme that uses temporal as well as spatial constraints and takes into consideration the local intensity characteristics of the image sequence. As a result, the algorithm creates image segments with smooth boundaries in both space and time without sacrificing spatial or temporal resolution. Thus, it provides image sequences with high spatial and temporal resolution and minimal gray-scale resolution. Similar approaches have been considered for segmentation of three-dimensional medical image data in [3] and [4]. However, there are significant differences between video sequences and three-dimensional still images. For computational efficiency as well as performance we use a multi-resolution approach.

### 2.1. Model

Let $\mathbf{y}$ be the three-dimensional (3-D) volume of images, and $y_t$ be the observed gray-scale image at time $t$. Each 2-D slice consists of a grid of sites $s$ and the intensity of a pixel at site $s$ is denoted by $y_{s,t}$, which typically takes values between 0 and 255. The pair $(s,t)$ can index each location in the 3-D volume. We model the spatio-temporal 3-D volume of the video sequence as a collection of regions of uniform or slowly varying intensity. Each region varies in shape and size and extends throughout the sequence of images. The only sharp transitions in gray level in space or time occur at the region boundaries.

A segmentation of the sequence of images into regions will be denoted by $\mathbf{x}$, where $x_t$ is the segmentation of the image $y_t$ into regions. Let $x_{s,t} = i$ mean that the pixel at $s$ in slice $y_t$ belongs to region $i$. The number of different region types (or classes) is $K = 2$. We model the segmentation distribution as a 3-D Gibbs random field.

The algorithm seeks to maximize the *a posteriori* probability density function $p(\mathbf{x}|\mathbf{y})$. By Bayes' theorem

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}) \tag{1}$$

where $p(\mathbf{x})$ is the *a priori* density of the distribution of regions and $p(\mathbf{y}|\mathbf{x})$ is the conditional density of the observed video sequence given the distribution of regions.

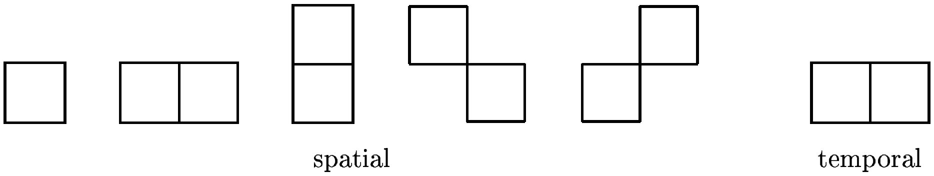


Figure 1: Clique types for Gibbs density

We model the distribution of regions by a 3-D Gibbs random field. We consider each image defined on the Cartesian grid and a neighborhood consisting of the 8 nearest pixels in the same 2-D slice and the two adjacent pixels at the identical site $s$ in the surrounding frames. The Gibbs density for $\mathbf{x}$ has the following form

$$p(\mathbf{x}) = \frac{1}{Z} \exp \left\{ - \sum_{C} V_C(\mathbf{x}) \right\} \tag{2}$$

where $Z$ is a normalizing constant, $V_C(\mathbf{x})$ are the clique potentials, and the summation is over all cliques $C$. A clique is a set of points that are neighbors of each other. A clique potential $V_C$ is a function that depends only on the pixels that belong to a clique $C$.

Our model assumes that the only nonzero potentials are those that correspond to the one- and two-point cliques shown in Fig. 1. In this simple case the cliques are either spatial (S) or temporal (T). The two-point clique potentials are defined as follows:

$$V_S(\mathbf{x}) = \begin{cases} -\beta_1, & \text{if } x_{s,t} = x_{q,t} \text{ and } (s,t),(q,t) \in S \\ +\beta_1, & \text{if } x_{s,t} \neq x_{q,t} \text{ and } (s,t),(q,t) \in S \end{cases} \tag{3}$$

$$V_T(\mathbf{x}) = \begin{cases} -\beta_2, & \text{if } x_{s,t} = x_{s,r} \text{ and } (s,t),(s,r) \in T \\ +\beta_2, & \text{if } x_{s,t} \neq x_{s,r} \text{ and } (s,t),(s,r) \in T \end{cases} \tag{4}$$

The parameters $\beta_1$ and $\beta_2$ are positive, so that two neighboring pixels are more likely to belong to the same class than to different classes. The clique potentials control the interaction between pixels within a single frame as well as across frames. We arbitrarily chose the total weight of the interaction within each frame to be equal to the total weight of the interaction across frames. Thus, $2\beta_2 = 8\beta_1$. We further assume that the one-point clique potentials are zero, which means that all region types are equally likely.

The conditional density is modeled as a white Gaussian process, with mean $\mu^i_{s,t}$ and variance $\sigma^2$. Each region $i$ is characterized by a different $\mu^i_{s,t}$ which is a slowly varying function of $s$ and $t$.

The combined probability density has the form

$$p(\mathbf{x}|\mathbf{y}) \propto \exp \left\{ - \sum_{t,s} \frac{1}{2\sigma^2} \left[ y_{s,t} - \mu^{x_{s,t}}_{s,t} \right]^2 - \sum_{C} V_C(\mathbf{x}) \right\} \tag{5}$$

We observe that the probability density function has two components. One constrains the region intensity to be close to the data; the other imposes spatial and temporal continuity. Since increasing $\sigma^2$ is equivalent to increasing $\beta_1$ and $\beta_2$, we fix $\beta_1$ and $\beta_2$ and estimate the noise variance $\sigma^2$. The noise variance controls the amount of detail in the segmentation. The performance of the algorithm is reasonable over a wide range of noise variances [2].

## 2.2. Algorithm

We now consider an iterative algorithm for estimating the distribution of regions $\mathbf{x}$ and the local intensity functions $\mu^i_{s,t}$ throughout the 3-D volume. At each frame $t$, the algorithm alternates between estimating $x_t$ and the intensity functions $\mu^i_{s,t}$. Note that the functions $\mu^i_{s,t}$ are defined on the same 3-D grid as the original gray-scale sequence $\mathbf{y}$ and the distribution of regions $\mathbf{x}$.

First, we consider the problem of estimating the local intensity functions in a frame at time $t$ (denoted by $\mu^i_t$). Given the region labels in the frame $x_t$ and the two surrounding frames $x_{t-1}$ and $x_{t+1}$, we estimate the intensity $\mu^i_{s,t}$ at each pixel $s$ in the frame by averaging the gray levels of all the pixels that belong to region $i$ and are inside a window of width $W$ centered at pixel $s$ in three consecutive frames.

The estimates of $\mu^i_{s,t}$ must be obtained for all region types $i$ and all pixels $s$ in each frame. To reduce computation, we obtain the estimates $\mu^i_{s,t}$ only on a grid of points in each frame, and use bilinear interpolation to obtain the remaining values. The spacing of the grid points in a frame is a function of the window size. We choose the spacing equal to half the window size in each spatial dimension (50% overlap). Since the functions $\mu^i_{s,t}$ are smooth, this is a good approximation. It also guarantees that the amount of computation is independent of window size.

Second, we consider the problem of estimating the distribution of regions. Given the intensity functions $\mu^i_{s,t}$, we use the Iterated Conditional Modes (ICM) approach proposed by Besag [5] to obtain an approximate MAP estimate for the distribution of regions $\mathbf{x}$.

The overall adaptive clustering algorithm alternates between estimating $\mathbf{x}$ and the intensity functions $\mu^i_{s,t}$. An initial estimate of $\mathbf{x}$ is obtained by the $K$-means algorithm [6] applied to each frame individually. In [1] we discussed a direct extension of the 1-D algorithm proposed in [2] as well as several suboptimal algorithms which offer advantages in terms of computational efficiency and time delay. Since computational efficiency and real time implementation were not critical in this project, we used Method A of [1] which offers the best performance for a reasonable amount of computation.

Method A [1] attempts to reduce the amount of computation while retaining the advantage of joint segmentation of all the frames in the 3-D volume. First, an initial estimate of $\mathbf{x}$ is obtained by the $K$-means algorithm applied to each frame individually. Given the region labels at some frame $x_t$ and the two surrounding frames $x_{t-1}$ and $x_{t+1}$, we estimate the intensity $\mu^i_{s,t}$ at each pixel $s$ in the frame. Then we update the estimate of $x_t$ using the ICM approach. The segmentation of the surrounding frames $x_{t-1}$ and $x_{t+1}$ is fixed. The algorithm then moves to the next frame and updates the estimates of $\mu^i_t$ and $x_t$, and so on, until all the frames are processed. Then the process is repeated. We define an *iteration* to consist of one update of $\mathbf{x}$ in all frames in the sequence. The window size for the intensity function estimation is kept constant until the procedure converges. The stopping criterion is that the update of $x_t$ at each frame in the volume converges in one cycle for all the frames. Weaker convergence criteria can be used to reduce the number of iterations. The whole procedure is then repeated with a smaller window size. The window depth (three consecutive frames) remains constant throughout the algorithm. The assumption is that scene characteristics remain fairly constant over time. A flowchart of the algorithm is given in Fig. 2. The algorithm stops when the minimum window size is reached. Typically we keep reducing the window size by a factor of two, until a minimum size of $W = 7$ pixels.

Finally, as in [2], we use a *multi-resolution* approach to improve algorithm performance and computational efficiency. First, we construct a pyramid of images at different resolutions. Then, the algorithm described above is performed on the images of the lowest resolution in the pyramid. When the minimum window size is reached, it moves to the next level in the pyramid and uses the current segmentation for all frames, expanded by two, as a starting point. As in [2], the starting window size for each level in the pyramid is twice the minimum window size of the previous level.

## 3. SUBJECTIVE EVALUATION TEST

The purpose of our experiment was to consider tradeoffs between temporal and gray-scale resolution for video conferencing and video phone applications. In particular, we were interested to find out for what frame rates of the full gray-scale sequence the observers prefer a binary sequence that preserves smooth motion and lip synchronization.

In the experiment we used an original gray-scale sequence at QCIF resolution, 30 frames/second, and 256 gray levels. The sequences shows a woman speaking on a technical subject. The total duration of the sequence was 48.9 secs (1467 frames). We obtained a two-level segmentation of this sequence at 30 frames/second using the algorithm described in the previous section. Figure 3 shows three consecutive frames of the gray-scale sequence and the corresponding segmentation. This two-level sequence was compared to the 256-level sequence at different frame rates: 1, 2, 3,

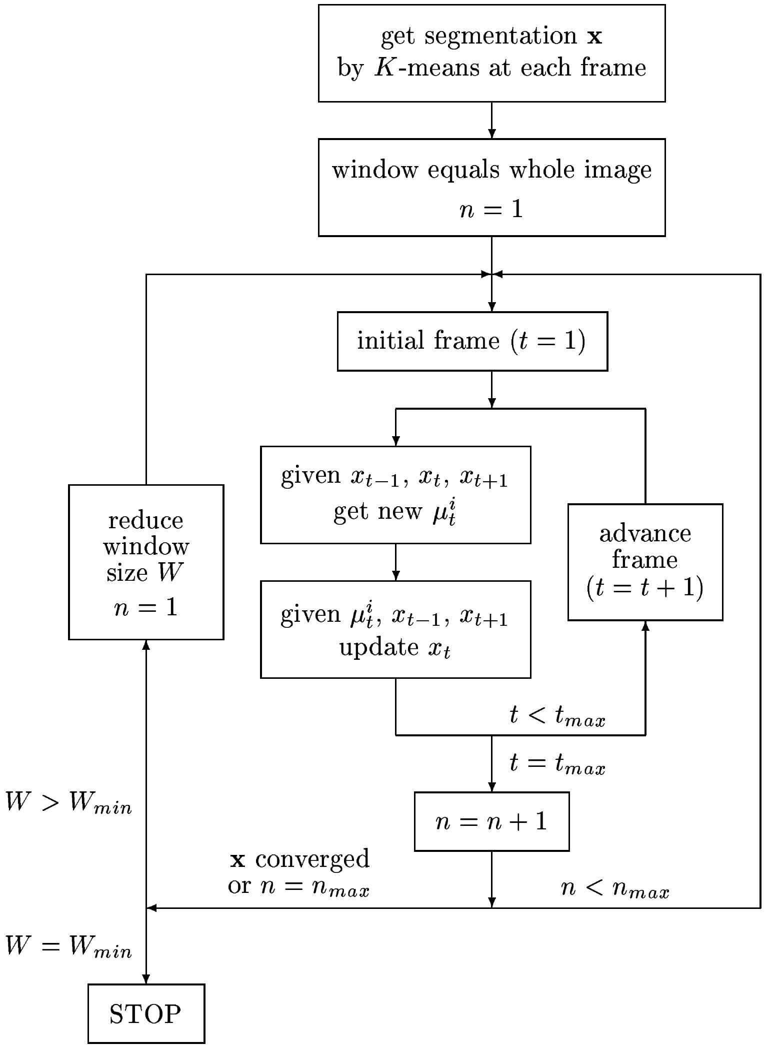

get segmentation $\mathbf{x}$ by $K$-means at each frame
window equals whole image $n = 1$
initial frame ($t = 1$)
given $x_{t-1}$, $x_t$, $x_{t+1}$ get new $\mu_t^i$
given $\mu_t^i$, $x_{t-1}$, $x_{t+1}$ update $x_t$
advance frame ($t = t + 1$)
reduce window size $W$ $n = 1$
$t < t_{max}$
$t = t_{max}$
$n = n + 1$
$W > W_{min}$
$\mathbf{x}$ converged or $n = n_{max}$
$n < n_{max}$
$W = W_{min}$
STOP

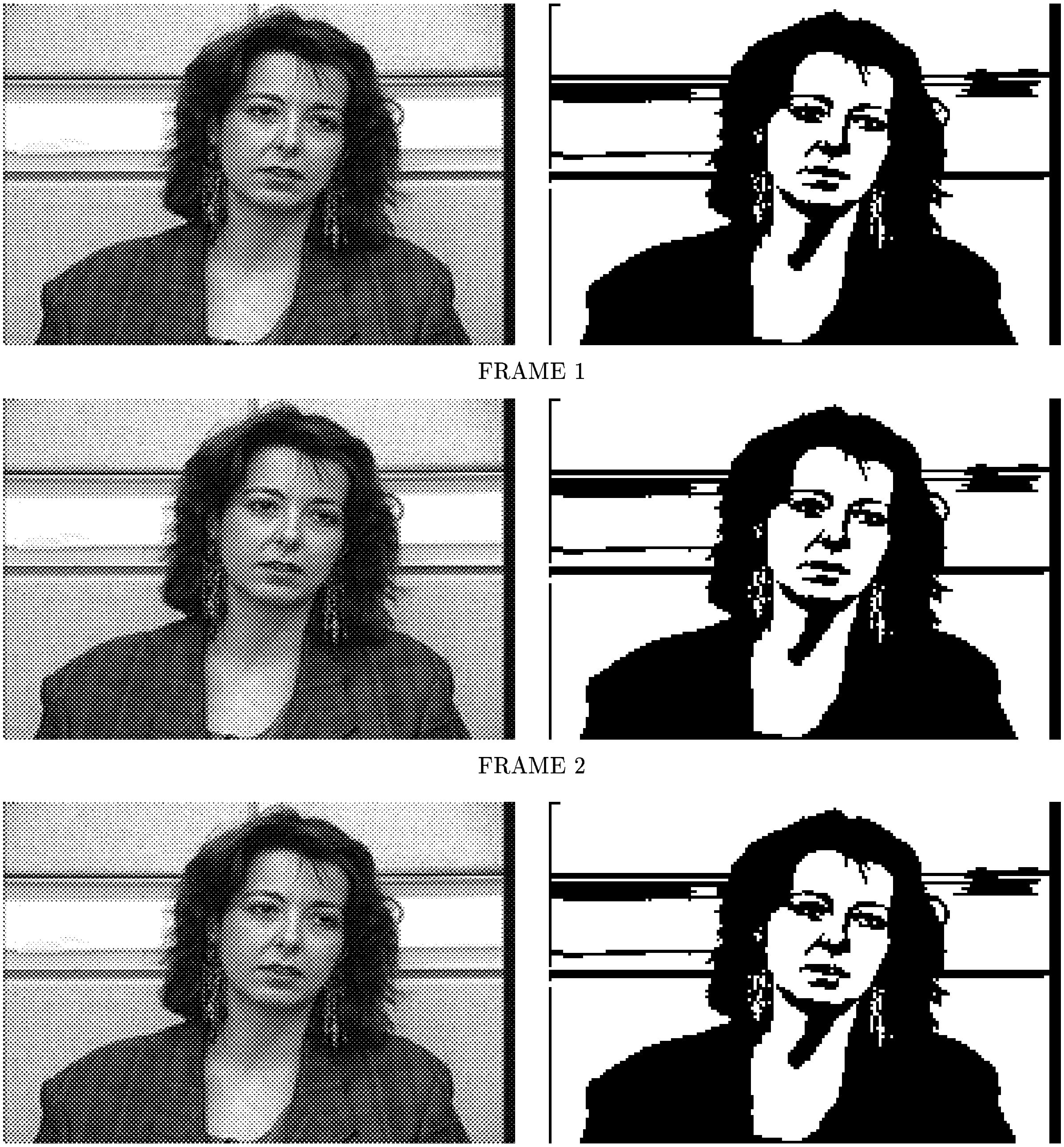
Figure 2: Adaptive clustering algorithm, Method A

3.75, 5, 6, and 7.5 frames/sec. The viewing distance was 20 image heights, the equivalent of two feet from an SGI workstation. The sequences were actually recorded on digital tape (D1) and played back on a large monitor.[2] The speech signal was not coded and of relatively high quality.

There were 23 participants in the subjective evaluation test. The test group included mostly people with technical backgrounds but also some nontechnical people. None of the observers had background in video compression. All were familiar with the speaker and comfortable with the topic she discusses.

Each observer was first shown the original gray-scale sequence, and was told to consider it as a standard against which to rate all other sequences. The observer was given the option to see the original again (none did) and was asked if he/she had any difficulty in understanding the speaker (none had). Then, the observer was shown seven different pairs of video sequences. Each pair consisted of the two-level sequence at 30 frames/sec and the gray-scale sequence at one of the above rates. The two sequences were shown one after the other, with the time between them fixed at 5 secs. The order of the pairs and the sequences within each pair was random. The time between different pairs was variable because the sequences were recorded on digital tape, and rewinding was necessary. After each pair was shown, the observer was asked to choose the sequence that he/she would prefer for communication. It was

[2] To avoid artifacts due to interlacing, the image frames were magnified by a factor of two using pixel repetition.



FRAME 1

FRAME 2

FRAME 3

Figure 3: Original gray-scale sequence and its two-level segmentation

explained that the two sequences differ from the original in a variety of ways and that the decision must be based on overall quality of perceived potential communication. The observers were not told that the audio signal was going to be the same for all sequences. At the end of the test, each observer was asked for comments and, in particular, what characteristics of the video influenced the decision and whether lip synchronization or eye contact was important. Each observer also completed a demographics survey to determine gender, vision, hearing, and familiarity with video communication systems.

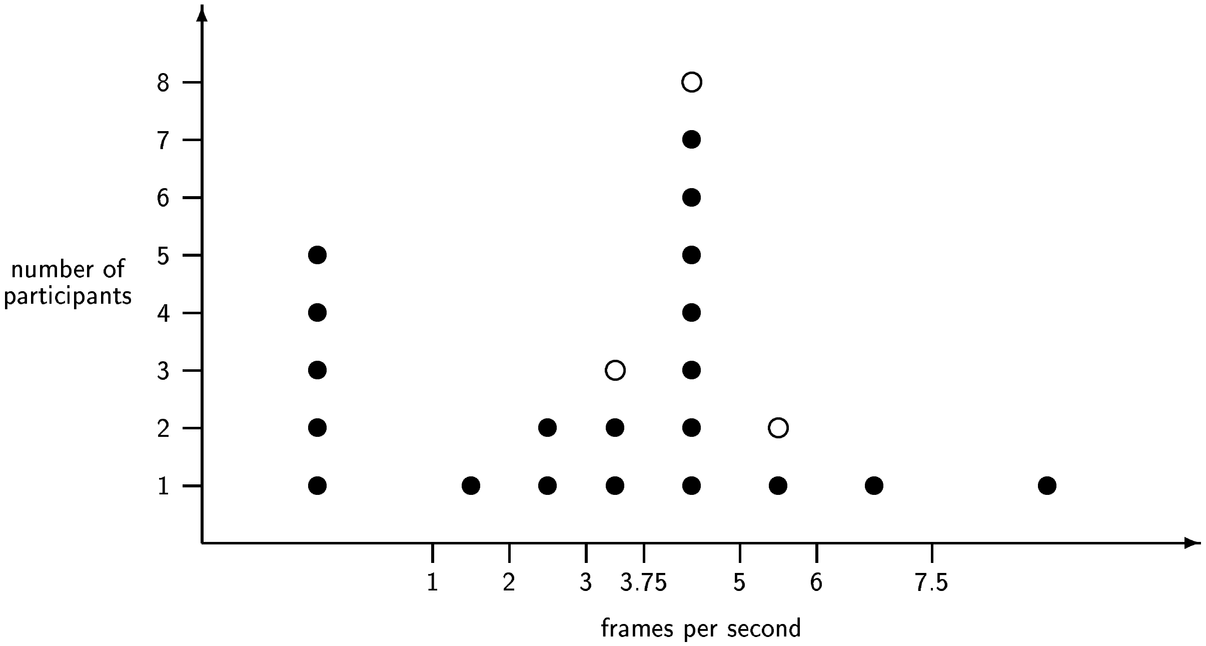


Figure 4: Test results

Since the gray-scale sequences were compared to a binary sketch, it was important to assume that the observers were familiar with the speaker. Indeed, all the test subjects were familiar with the woman in the video sequence. It was made clear that the purpose of the video sequence was to communicate with someone you already know, not to meet a person you have never seen before.

The test results are shown in Figure 4. The horizontal axis shows the different frame rates for the gray-scale sequence. Each observer is denoted by a black dot. For all the rates to the right of a black dot, the observer chose the gray-scale sequence over the two-level sequence. For all the rates to the left of the dot, the observer chose the two-level sequence over the gray-scale sequence. The hollow dots indicate special cases, discussed below.

The test results indicate that, when the frame rate of the full gray-scale sequence is low, most observers prefer the binary sequence that preserves smooth motion and lip synchronization. Most observers preferred the binary sequence when the rate of the gray-scale sequence was less than 5 frames/sec. About 25% of the observers prefer the "clear" (full gray-scale) images to the binary sketches at all rates, indicating that the tradeoff we chose was a little too extreme. Actually, this strengthens the results because it indicates that the two-level sequence is quite objectionable. Therefore, one would choose it only if the gray-scale sequence at a low frame rate is really annoying. One observer chose the gray-scale sequence at high rates (6 frames/sec and above, shown with a hollow dot) and also at low rates (1 frame/sec), and chose the two-level sequence at the intermediate rates. That is, according to this observer, the moving sequence should either be synchronized with the speech, or it should be a sequence of almost still images. Two of the observers, also shown with hollow dots, changed their mind during the test. The final choice is shown in the figure.

Most of the observers (12) indicated that lip synchronization was important in their decision and a few (2) mentioned that jerkiness was annoying. One observer pointed out that the binary sketches showed some expressions because of the movement. The observers that chose the gray-scale sequence at all rates indicated that it is important to see a "reasonable" or "clear" picture of the person, that they wanted to see the expression of the face and to get feedback about what the other person is feeling, and that the artifacts of the two-level sequence were unacceptable. Finally, one observer chose the two-level sequence at all rates and indicated that, since the audio was very clear, the video was not needed at all.

If we exclude the observers that found the binary sketches unacceptable, the distribution of the remaining observers is a nice bell-shaped curve that indicates that 5 frames/sec is probably a critical rate for video conferencing. The jerkiness and lack of lip synchronization at rates lower than 5 frames/sec is very objectionable and people are ready to make significant compromises to avoid it.

## 4. CONCLUSIONS

We considered tradeoffs between temporal and gray-scale resolution for video conferencing and video phone applications. We conducted a subjective evaluation test to determine how the quality of sequences with different gray-scale and temporal resolutions is perceived in the presence of an audio signal. In particular, we explored the importance of the synchronization of lip movements with speech.

The test results indicate that, when the frame rate of the full gray-scale sequence is low (less than 5 frames/sec), most observers prefer a two-level sequence that preserves smooth motion and lip synchronization. The tradeoffs considered in this paper are intended to help in the design of a video compression scheme and the choice of a display device.

## 5. ACKNOWLEDGMENTS

The authors would like to thank Darren Kall for his help in designing and debugging the experiment. His comments and suggestions were invaluable. We would also like to thank James Pawlyk for his help with the video equipment and, in particular, with the integration of audio and video. Finally, we would like to thank all the participants in the experiment.

## 6. REFERENCES

[1] R. O. Hinds and T. N. Pappas, "An adaptive clustering algorithm for segmentation of video sequences," in *Proc. ICASSP-95*, (Detroit, MI), May 1995.

[2] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Proc.*, vol. SP-40, pp. 901–914, Apr. 1992.

[3] M. M. Chang, A. M. Tekalp, and M. I. Sezan, "Bayesian segmentation of MR images using 3-D Gibbsian priors," in *Proc. SPIE, vol. 1903, Image and Video Processing*, pp. 122–133, 1993.

[4] C. W. Chen, J. Luo, K. J. Parker, and T. S. Huang, "3D image segmentation via adaptive K-mean clustering and knowledge-based morphological operations," *IEEE Trans. Image Proc.* to appear.

[5] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

[6] R. M. Gray and Y. Linde, "Vector quantizers and predictive quantizers for Gauss-Markov sources," *IEEE Trans. Comm.*, vol. COM-30, pp. 381–389, Feb. 1982.