Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 478396, 10 pages
doi:10.1155/2008/478396

Research Article

An Iterative Decoding Algorithm for Fusion of Multimodal Information

Shankar T. Shivappa, Bhaskar D. Rao, and Mohan M. Trivedi

Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Correspondence should be addressed to Shankar T. Shivappa, sshivappa@ucsd.edu

Received 16 February 2007; Revised 19 July 2007; Accepted 26 October 2007

Recommended by Eric Pauwels

Human activity analysis in an intelligent space is typically based on multimodal informational cues. Use of multiple modalities gives us a lot of advantages. But information fusion from different sources is a problem that has to be addressed. In this paper, we propose an iterative algorithm to fuse information from multimodal sources. We draw inspiration from the theory of turbo codes. We draw an analogy between the redundant parity bits of the constituent codes of a turbo code and the information from different sensors in a multimodal system. A hidden Markov model is used to model the sequence of observations of individual modalities. The decoded state likelihoods from one modality are used as additional information in decoding the states of the other modalities. This procedure is repeated until a certain convergence criterion is met. The resulting iterative algorithm is shown to have lower error rates than the individual models alone. The algorithm is then applied to a real-world problem of speech segmentation using audio and visual cues.

1. INTELLIGENT SPACES AND MULTIMODAL SYSTEMS

Intelligent environments facilitate a natural and efficient mechanism for human-computer interaction and human activity analysis. An intelligent space can be any physical space that possesses the following requirements [1].

(i) Intelligent spaces should facilitate normal human activities taking place in these spaces.
(ii) Intelligent spaces should automatically capture and maintain awareness of the events and activities taking place in these spaces.
(iii) Intelligent spaces should be responsive to specific events and triggers.
(iv) Intelligent spaces should be robust and adaptive to various dynamic changes.

An intelligent space can be a room in a building or an outdoor environment. Designing algorithms for such spaces involves the real-world challenges of real-time, reliable, and robust performance over the wide range of events and activities, which can occur in these spaces. In this paper, we consider an indoor meeting room scenario. Though the high-level framework that is presented below can be applied in other scenarios, we have chosen to restrict our initial investigations to a meeting room. Much of the framework in this case has been summarized in [2].

Intelligent spaces have sensor-based systems that allow for natural and efficient human-computer interaction. In order to achieve this goal, the intelligent space needs to analyze the events that take place and maintain situational awareness. In order to analyze such events automatically, it is essential to develop mathematical models for representing different kinds of events and activities.

Research efforts in the field of human activity analysis have increasingly come to rely on multimodal sensors. Analyzing multimodal signals is a necessity in most scenarios and has added advantages in others [1, 3].

Human activity is essentially multimodal. Voice and gesture, for example, are intimately connected [4]. Researchers in automatic speech recognition (ASR) have used the multimodal nature of human speech to enhance the ASR accuracy and robustness [5]. Recent research efforts in human activity analysis have increasingly come to include multimodal sensors [6–8].

Certain tasks that are very difficult to handle with unimodal sensors might become tractable with the use of multimodal sensors. The limitations of audio analysis in reverberant environments have been discussed in [9]. But the addition of the video modality can solve problems like source localization in reverberant environments. In fact, video analysis can even provide a more detailed description of the subject's state-like emotion, and so forth, as shown in [10]. But the use of video alone has some disadvantages as seen in [11]. Even in the simple task of speech segmentation, it is conceivable that some nasal sounds can be produced without any movement of the mouth, and conversely movement of the mouth alone, like in yawning, might not signify the presence of speech. By combining the strengths of each modality, a multimodal solution for a set of tasks might be simpler than putting together the unimodal counterparts.

Multiple modalities carry redundant information on complimentary channels, and hence they provide robustness to environmental and sensor noises that might affect each of these channels differently. Due to these reasons, we focus our attention towards building multimodal systems for human activity analysis.

1.1. Fusion of information in multimodal systems

Fusion of information from different streams is a big challenge in multimodal systems. So far, there has not been any standard fusion technique that has been widely accepted in the published literature. Graphical models have been widely discussed as the most suitable candidates for modeling and fusion information in multimodal systems [12].

Information fusion can occur at various levels of a multimodal system. A sensor-level fusion of video signals from normal and infrared cameras is used for stereo analysis in [13]. At a higher level is the feature-level fusion. The audio and visual features used together in the ASR system built at John Hopkins University, 2000 workshop [5], are a good example of feature-level fusion. Fusions at higher levels of abstraction (decision level) have also been proposed. Graphical models have been frequently used for this task [12]. Fusion at the sensor level is appropriate when the modalities to be fused are similar. As we proceed to the feature level, fusing more disparate sources becomes possible. At the decision level, all the information is represented in the form of probabilities, and hence it is possible to fuse information from a wide variety of sensors. In this paper, we develop a general fusion algorithm at the decision level.

In this paper, we develop a fusion technique in the hidden Markov model (HMM) framework. HMMs are a class of graphical models that have been used traditionally in speech recognition and human activity analysis [14]. We plan to extend our algorithm to more general graphical models in the future. Our scheme uses HMMs trained on unimodal data and merges the decisions (a posteriori probabilities) from different modalities. Our fusion algorithm is motivated by the theory of iterative decoding.

1.2. Advantages of the iterative decoding scheme

A good fusion scheme should have lower error rates than those obtained from the unimodal models. Both the joint modeling framework and the iterative decoding framework have this property. Multimodal training data is hard to obtain. Iterative decoding overcomes this problem by utilizing models trained on unimodal data. Building joint models on the other hand requires significantly greater amounts of multimodal data than training unimodal models due to the increase in dimensionality or complexity of the joint model or both. Working with unimodal models also makes it possible to use a well-learned model in one modality to segment and generate training data for the other modalities, thus overcoming the problem of the lack of training data to a great extent.

In many applications like ASR, well-trained unimodal models might already be available. Iterative decoding utilizes such models directly. Thus, extending the already existing unimodal systems to multimodal ones is easier. Another common scheme used to integrate unimodal HMMs is the product HMM [15]. In our simulations, we see that the product rule performs as well as the joint model. But the product rule has the added disadvantage that it assumes a one-to-one correspondence between the hidden states of the two modalities. The generalized multimodal version of the iterative decoding algorithm (see Section 5) relaxes this requirement. Moreover, the iterative decoding algorithm performs better than the joint model and the product HMM in the presence of background noise, even in cases where there is a one-to-one correspondence between the two modalities.

In noisy environments, the frames affected by noise in different modalities are at best nonoverlapping and at worst independent. The joint models are not able to separate out the noisy modalities from the clean ones. Because of this reason, the iterative decoding algorithm outperforms the joint model at low SNR. In case of other decision-level fusion algorithms like the multistream HMMs [16] and reliability-weighted summation rule [17], one has to estimate the quality (SNR) of the individual modalities to obtain good performance. Iterative decoding does not need such a priori information. This is a very significant advantage of the iterative decoding scheme because the quality of the modalities is in general time-varying. For example, if the speaker keeps turning away from the camera, video features are very unreliable for speech segmentation. The exponential weighting scheme of multistream HMMs requires real-time monitoring of the quality of the modalities which in itself is a very complex problem.

2. TURBO CODES AND ITERATIVE DECODING

Turbo codes are a class of convolutional codes that perform close to the Shannon limit of channel capacity. The seminal paper by Berrou et al. [18] introduced the concept of iterative decoding to the field of channel coding. Turbo codes achieve their high performance by using two simple codes, working in parallel to achieve the performance of single complex code. The iterative decoding scheme is a method to

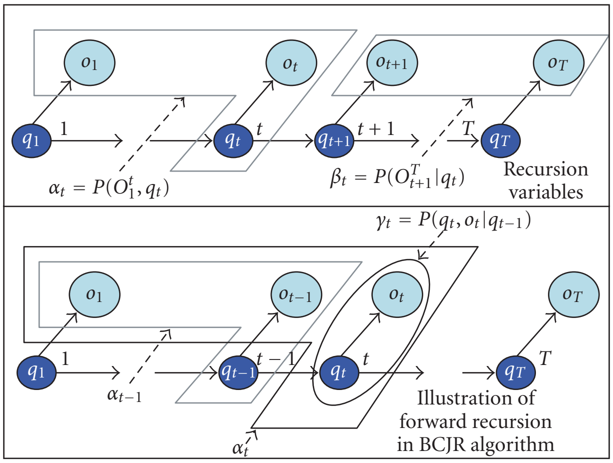

Figure 1: Illustrating the forward recursion of the BCJR algorithm.

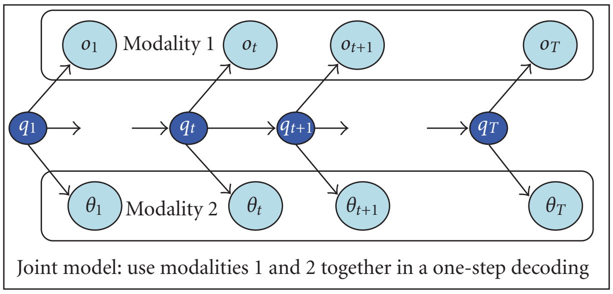

Figure 2: Joint model for a bimodal scenario.

combine the decisions from the two decoders at the receiver and achieve high performance. In other words, two simple codes working in parallel perform as well as a highly complex code which in practice cannot be used due to complexity issues.

We draw an analogy between the redundant information of the two channels of a turbo code and the redundant information in the multiple modalities of a multimodal system. We develop a modified version of the iterative decoding algorithm to extract and fuse the information from parallel streams of multimodal data.

3. FORMALIZATION OF THE PROBLEM

Let us consider a multimodal system to recognize certain patterns of activity in an intelligent space [1]. It consists of multimodal sensors at the fundamental level. From the signals captured by these sensors, we extract feature vectors that encapsulate the information contained in the signals in finite dimensions. Once the features are selected, we model the activity to be recognized, statistically. For an activity that involves temporal variation, hidden Markov models (HMMs) are a popular modeling framework [14].

3.1. *Hidden Markov models*

Let $\lambda = (A, \pi, B)$ represent the parameters of an HMM with N hidden states, that model a particular activity. Now, the decoding problem is to estimate the optimal state sequence $Q_1^T = \{q_1, q_2, \ldots, q_T\}$ of the HMM based on the sequence of observations $O_1^T = \{o_1, o_2, \ldots, o_T\}$.

The maximum a posteriori probability state sequence is provided by the BCJR algorithm [19]. The MAP estimate for the hidden state at time t is given by $\hat{q}_t = \arg\max P(q_t, O_1^T)$. The BCJR algorithm computes this using the forward (see Figure 1) and backward recursions.

Define

$$\begin{gathered}
\lambda_t(m) = P(q_t = m, O_1^T), \\
\alpha_t(m) = P(q_t = m, O_1^t), \\
\beta_t(m) = P(O_{t+1}^T \mid q_t = m), \\
\gamma_t(m', m) = P(q_t = m, o_t \mid q_{t-1} = m'), \\
m = 1, 2, \ldots, N, \quad m' = 1, 2, \ldots, N.
\end{gathered} \tag{1}$$

Then establish the recursions

$$\begin{aligned}
\alpha_t(m) &= \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m), \\
\beta_t(m) &= \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m'), \\
\lambda_t(m) &= \alpha_t(m) \cdot \beta_t(m).
\end{aligned} \tag{2}$$

These enable us to solve for the MAP state sequence given appropriate initial conditions for $\alpha_1(m)$ and $\beta_T(m)$.

3.2. *Multimodal scenario*

For the sake of clarity, let us consider a bimodal system. There are observations O_1^T from one modality and observations $\Theta_1^T = \{\theta_1, \theta_2, \ldots, \theta_T\}$ from the other modality. The MAP solution in this case would be $\hat{q}_t = \arg\max P(q_t, O_1^T, \Theta_1^T)$. In order to apply the BCJR algorithm to this case, we can concatenate the observations (feature-level fusion) and train a new HMM in the joint feature space. Instead of building a joint model, we develop an iterative decoding algorithm that allows us to approach the performance of the joint model by iteratively exchanging information between the simpler models and updating their posterior probabilities.

4. ITERATIVE DECODING ALGORITHM

This is a direct application of the turbo decoding algorithm [18]. In this section, it is assumed that the hidden states in the two modalities have a one-to-one correspondence. This requirement is relaxed in the generalized solution presented in the next section.

In the first iteration of the iterative algorithm, we decode the hidden states of the HMM using the observations from the first modality, O_1^T. We obtain the a posteriori probabilities $\lambda_t^{(1)}(m) = P(q_t = m, O_1^T)$.

In the second iteration, these a posteriori probabilities, $\lambda_t^{(1)}(m)$, are utilized as extrinsic information in decoding the

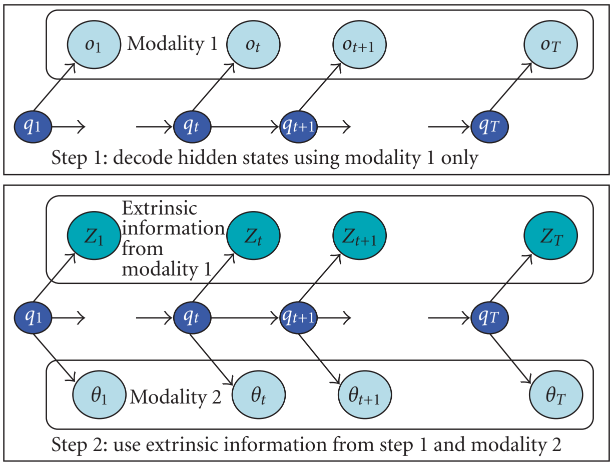

FIGURE 3: First two steps of the iterative decoding algorithm.

hidden states from the observations of the second modality Θ_1^T (see Figure 3). Thus the a posteriori probabilities in the second stage of decoding are given by $\lambda_t^{(2)}(m) = P(q_t = m, \Theta_1^T, Z^{(1)}{}_1^T)$, where $Z_t^{(1)} = \lambda_t^{(1)}$ is the extrinsic information from the first iteration.

4.1. Modified BCJR algorithm for incorporating the extrinsic information

In order to evaluate $\lambda_t^{(2)}$, we modify the BCJR algorithm as follows:

$$\begin{aligned}
\lambda_t^{(2)}(m) &= P(q_t = m, \Theta_1^T, Z^{(1)}{}_1^T), \\
\alpha_t^{(2)}(m) &= P(q_t = m, \Theta_1^t, Z^{(1)}{}_1^t), \\
\beta_t^{(2)}(m) &= P(\Theta_{t+1}^T, Z^{(1)}{}_{t+1}^T \mid q_t = m), \\
\gamma_t^{(2)}(m', m) &= P(q_t = m, \theta_t, Z_t^{(1)} \mid q_{t-1} = m').
\end{aligned} \tag{3}$$

Then the recursions do not change, except for the computation of $\gamma_t^{(2)}(m', m)$.

Since the extrinsic information is independent of the observations from the second modality, $\gamma_t^{(2)}(m', m) = P(q_t = m \mid q_{t-1} = m') \cdot P(\theta_t \mid q_t = m) \cdot P(Z_t^{(1)} \mid q_t = m)$.

Here $Z_t^{(1)} = [z_{1t}^{(1)}, z_{2t}^{(1)}, \ldots, z_{Nt}^{(1)}]'$ is a vector of probability values. A histogram of each component of $Z_t^{(1)}$ for $q_t = 2$ in an $N = 4$ state HMM synthetic problem is shown in Figure 4. From the histogram, one can see that a simple parametric probability model for $P(Z_t^{(1)} \mid q_t = m)$ is obtained as

$$P(Z_t^{(1)} \mid q_t = m) = f(1 - z_{mt}^{(1)}; \rho) \cdot \prod_{i \neq m} f(z_{it}^{(1)}; \rho), \tag{4}$$

where

$$f(x; \rho) = \begin{cases} \dfrac{1}{\rho} e^{-x/\rho}, & x \geq 0, \\ 0, & x < 0 \end{cases} \tag{5}$$

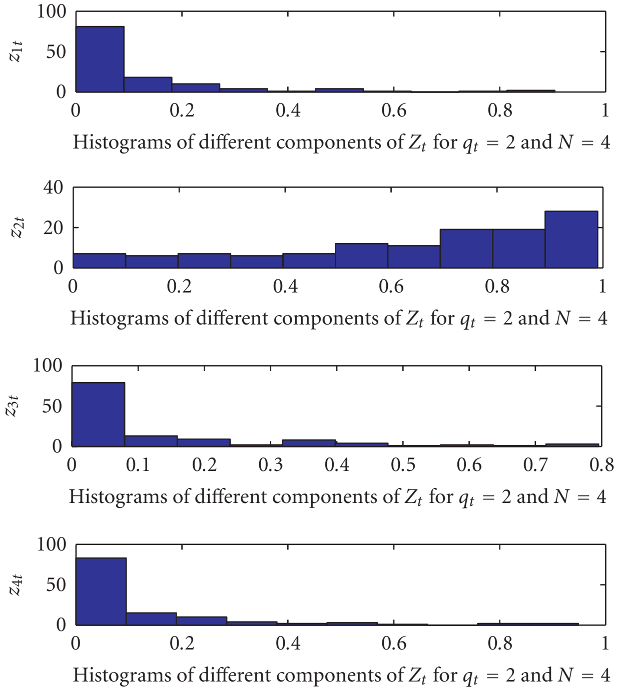

FIGURE 4: A histogram of each component of Z_t for $q_t = 2$ in an $N = 4$ state HMM synthetic problem.

is an exponential distribution with rate parameter $1/\rho$. Other distributions like the beta distribution could also be used. The exponential distribution is chosen due to its simplicity.

In the third iteration, the extrinsic information to be passed back to decoder 1 is the a posteriori probabilities $\lambda_t^{(2)}(m)$. But part of this information, $(\lambda_t^{(1)}(m))$, came from decoder 1 itself. If we were to use $\lambda_t^{(2)}$ as the extrinsic information in the third iteration, it would destroy the independence between the observations from the first modality and the extrinsic information. We overcome this difficulty by choosing another formulation for the extrinsic information based on the following observation:

$$\begin{aligned}
\lambda_t^{(2)}(m) &= \alpha_t^{(2)}(m) \cdot \beta_t^{(2)}(m), \\
\alpha_t^{(2)}(m) &= \sum_{m'} \alpha_{t-1}^{(2)}(m') \cdot \gamma_t^{(2)}(m', m), \\
\lambda_t^{(2)}(m) &= \sum_{m'} \alpha_{t-1}^{(2)}(m') \cdot \gamma_t^{(2)}(m', m) \cdot \beta_t^{(2)}(m), \\
\lambda_t^{(2)}(m) &= P(Z_t^{(1)} \mid q_t = m) \sum_{m'} \alpha_{t-1}^{(2)}(m') \\
&\quad \cdot P(q_t = m \mid q_{t-1} = m') \cdot P(\theta_t \mid q_t = m) \cdot \beta_t^{(2)}(m), \\
\lambda_t^{(2)}(m) &= P(Z_t^{(1)} \mid q_t = m) \cdot Y_t^{(2)}.
\end{aligned} \tag{6}$$

Note that $Y_t^{(2)}$ does not depend on $Z_t^{(1)}$ and it is hence uncorrelated with o_t. This argument follows the same principles used in turbo coding literature [18]. Hence, we normalize

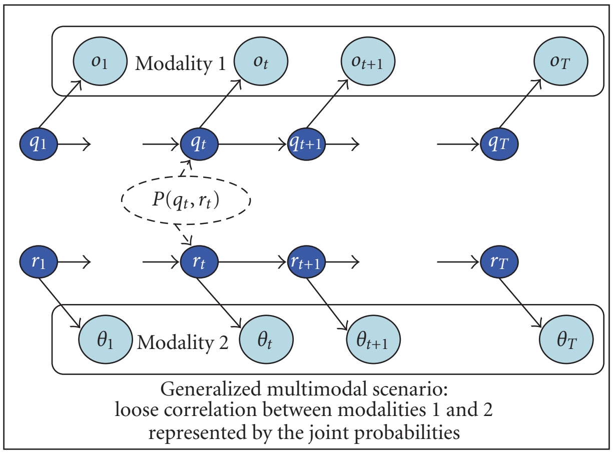

Figure 5: A more generalized bimodal problem.

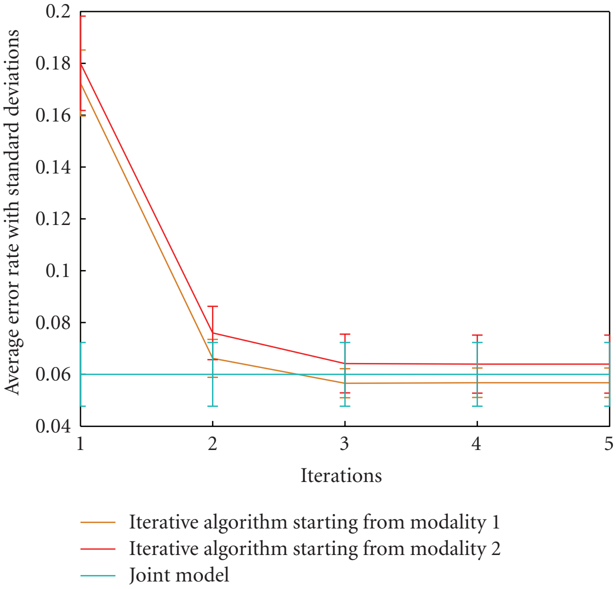

Figure 6: Error rate at different iterations for a 4-state HMM problem with one-to-one correspondence between the two modalities. Note the convergence of the error rate to that of the joint model.

$Y_t^{(2)}$ to sum to 1 and consider the normalized vector to be the extrinsic information passed on to decoder 1 in the third iteration.

The normalized extrinsic information $Z_t^{(2)}(m) = \lambda_t^{(2)}(m) / P(Z_t^{(1)} \mid q_t = m) / \sum_{m'} \lambda_t^{(2)}(m') / P(Z_t^{(1)} \mid q_t = m')$ is passed back to decoder 1.

The iterations are continued till the state sequences converge in both modalities or a fixed number of iterations are reached.

5. GENERAL MULTIMODAL PROBLEM

In the previous section, we assumed that the hidden states in the two modalities of a multimodal system are the same. In this section, we loosen this restriction and allow the hidden states in the individual modalities to just have a known prior co-occurrence probability (see Figure 5). In particular, if q_t and r_t represent the hidden states in modalities 1 and 2 at time t, then we know the joint probability distribution $P(q_t = m, r_t = m')$ and assume this to be stationary.

This corresponds to the case where there is a loose but definite interaction between the two modalities as seen very clearly in the case of phonemes and visemes, in audiovisual speech recognition. There is no one-to-one correspondence between visemes and phonemes. But the occurrence of one phoneme corresponds to the occurrence of a few specific visemes and vice versa.

5.1. *Iterative decoding algorithm in the general case*

This is an extension of the iterative decoding algorithm as presented in the turbo coding scenario. In this case, we have the same steps as in the iterative algorithm of Section 4. But at the jth iteration in the modified BCJR algorithm, in the computation of $\gamma_t^{(j)}(m', m) = P(q_t = m, \theta_t, Z_t^{(j-1)} \mid q_{t-1} = m')$, we now need to compute

$$\begin{aligned}
\gamma_t^{(j)}(m', m) &= P(r_t = m, \theta_t, Z_t^{(j-1)} \mid r_{t-1} = m'),\\
\gamma_t^{(j)}(m', m) &= P(r_t = m \mid r_{t-1} = m') \cdot P(\theta_t \mid r_t = m)\\
&\qquad \cdot P(Z_t^{(j-1)} \mid r_t = m),\\
\gamma_t^{(j)}(m', m) &= P(r_t = m \mid r_{t-1} = m') \cdot P(\theta_t \mid r_t = m)\\
&\qquad \cdot \sum_n \{P(Z_t^{(j-1)} \mid q_t = n) P(q_t = n \mid r_t = m)\},
\end{aligned} \tag{7}$$

which can be computed by the joint probability distribution $P(q_t = m, r_t = m')$.

The rest of the iterative algorithm remains the same as before.

6. EXPERIMENTAL VERIFICATION OF THE ITERATIVE DECODING ALGORITHM

In this section, we present the results of applying the iterative decoding algorithm to a synthetic problem. We choose a synthetic problem in order to validate our algorithm before applying it to a real-world problem so as to isolate the performance characteristics of our algorithm from the complexities of real-world data, which are dealt with in Section 7.

We generate observations from an HMM with 4 states whose observation densities are 4-dimensional Gaussian distributions. We also construct a joint model by concatenating the feature vectors. The goal of the experiment is to decode the state sequence from the observations and compare it with the true state sequence in order to obtain the error rates. The experiment is repeated several times and the average error rates are obtained.

In the first case, the joint model with 8 dimensions and 4 states was used to generate the state and observation sequences. We use the joint model to decode the state sequence from the observations. Next, we consider the observations

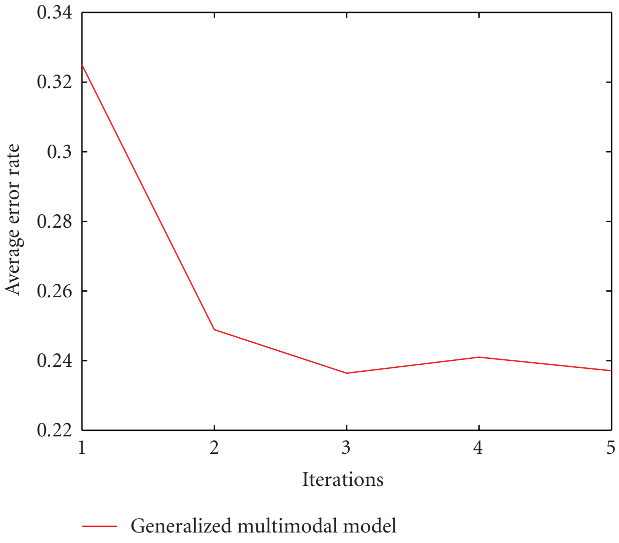

Figure 7: Error rate at different iterations for a generalized multimodal problem. Note that the performance follows the same trend as in the previous case.

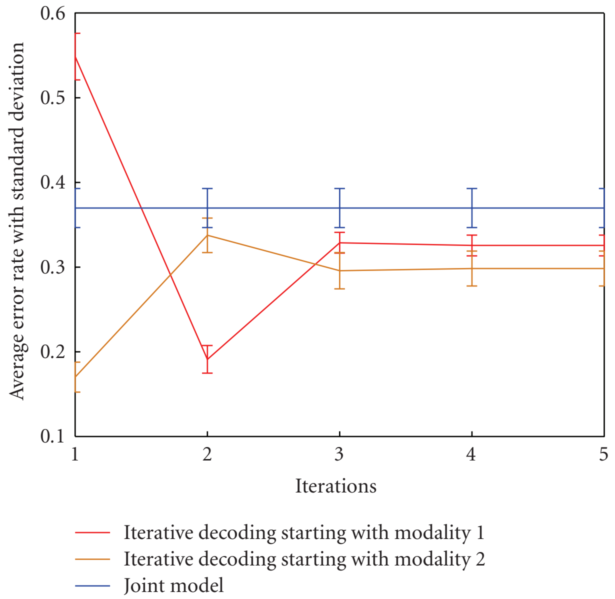

Figure 8: Error rate at different iterations in the case of noisy modalities. Note that the iterative algorithm performs better than the joint model at low SNR.

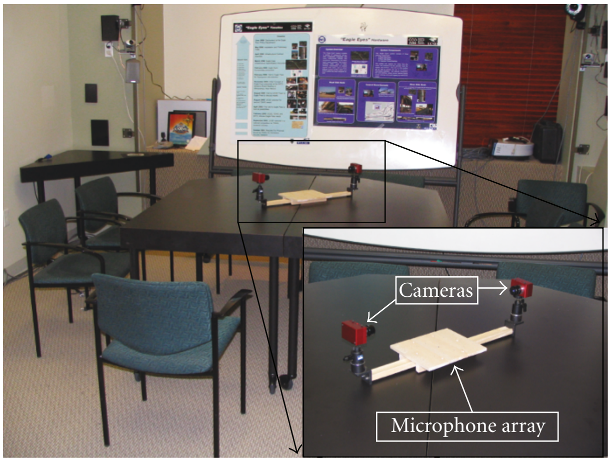

Figure 9: Testbed and the associated audio and video sensors.

to be generated by two modalities with 4 dimensions each. We now consider the product rule [15] as another alternative to the joint model. But in our simulations, we found its error rates to be the same as those of the joint model. Hence we assume the joint model to give us the baseline performance. The iterative decoding algorithm described in Section 4 is applied to decode the state sequence and compare it with the true state sequence. The results are plotted in Figure 6. We can see that the iterative decoding algorithm converges to the baseline performance and it reduces the error rate by almost 50% compared to the unimodal case (iteration 1). Figure 6 also shows the standard deviation of error from which it can be seen that the performance is indeed close to the baseline performance. Since the two modalities have similar unimodal error rates, the error dynamics of the iterative algorithm are independent of the starting modality.

In the second example, we generate observations from two independent HMMs such that the state sequence follows a known joint distribution. We then apply the generalized iterative decoding algorithm described in Section 5. The results are shown in Figure 7. In this case, we do not have a baseline experiment for comparison as the two streams are only loosely coupled, but the general trend in average error rate with each iteration is similar to the case shown in Figure 6.

In the presence of noise, the iterative algorithm outperforms the joint model as shown in Figure 8. Based on the standard deviation of error, a standard *t*-test reveals that the difference between the joint model and the iterative decoding algorithm is statistically significant after the third iteration. In this case, we added additive white Gaussian noise to the features of one of the modalities. No a priori information about the noise statistics is assumed to be available. Note that in this case, the individual modalities have varying noise levels, and hence the convergence of the iterative algorithm is dependent on the starting modality. But in both cases, we see that the iterative algorithm converges to the same performance after the third iteration. This illustrates the advantage of iterative decoding over joint modeling as mentioned in Section 1.1.

7. EXPERIMENTAL TESTBED

In this section, we describe an experimental testbed that is set up at the Computer Vision and Robotics Research (CVRR) lab at the University of California, San Diego. The goal of this exercise is to develop and evaluate human activity analysis algorithms in a meeting room scenario. Figure 9 shows a detailed view of the sensors deployed.

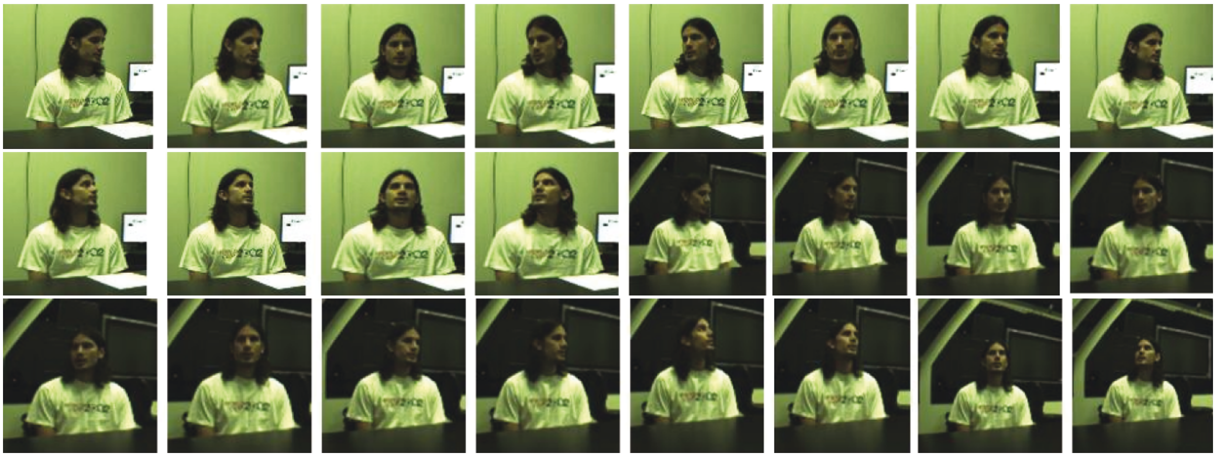

Figure 10: Different head poses and backgrounds for one subject out of 20 subjects in our database.

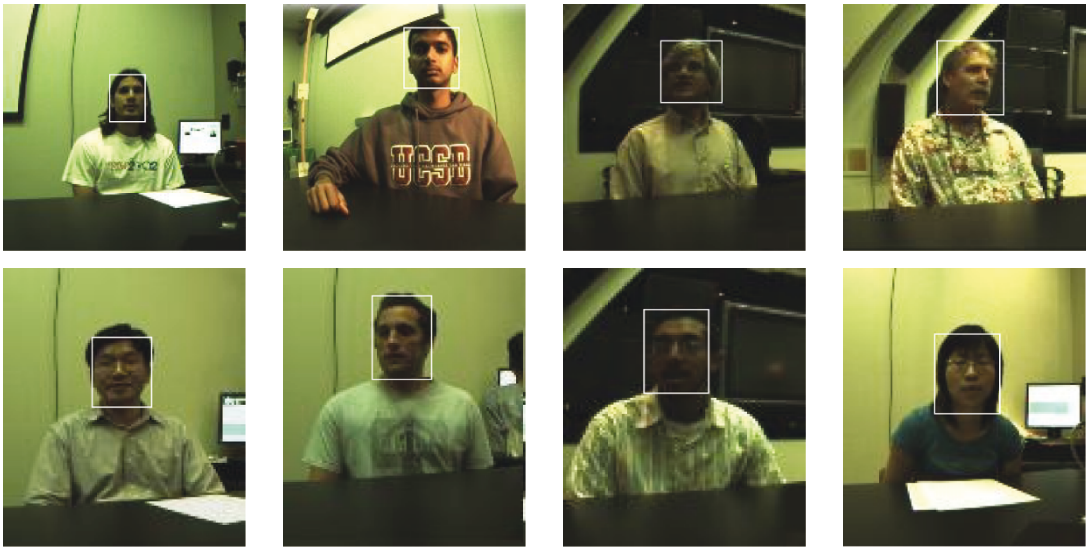

Figure 11: Face detection using the Viola-Jones face detector with various subjects.

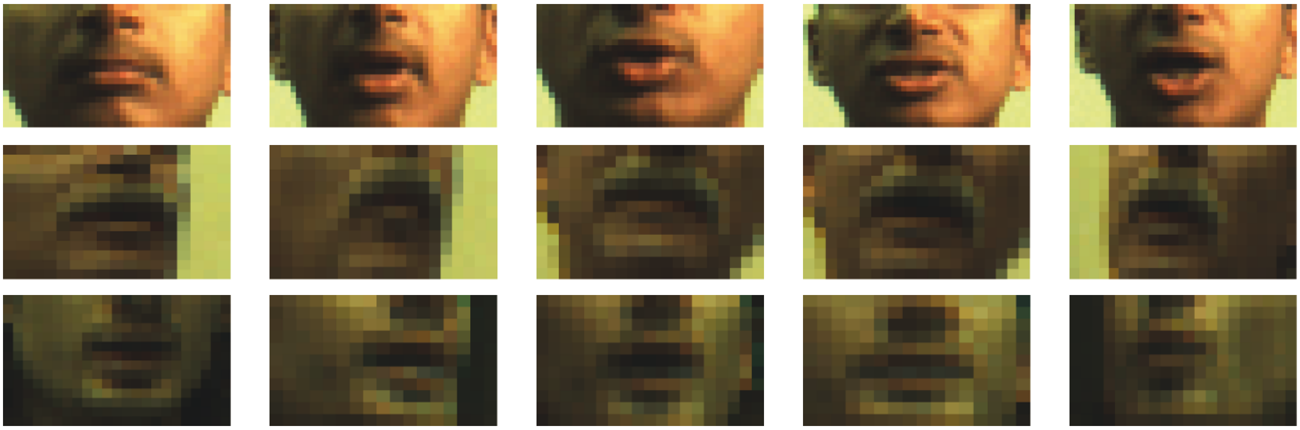

Figure 12: Some snapshots of the lip region during a typical utterance. Observe the variations in pose and facial characteristics of the three different subjects, which limit the performance of a video-only system.

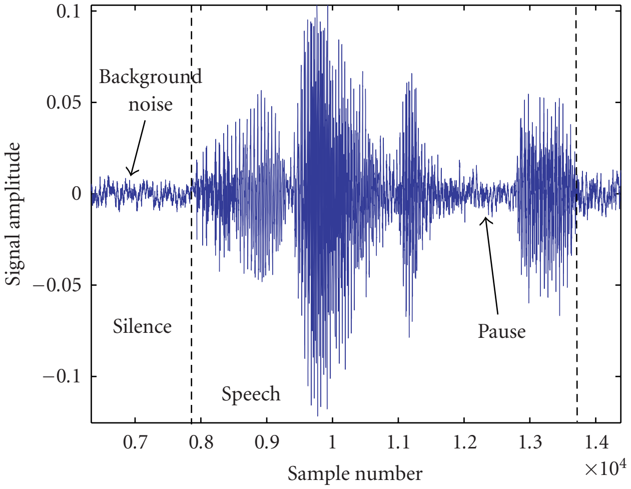

FIGURE 13: Audio waveform of speech in background noise. The short pauses between words which can be confused by an audio-only system for background noise will be detected as speech by the video modality, based on the lip movement.

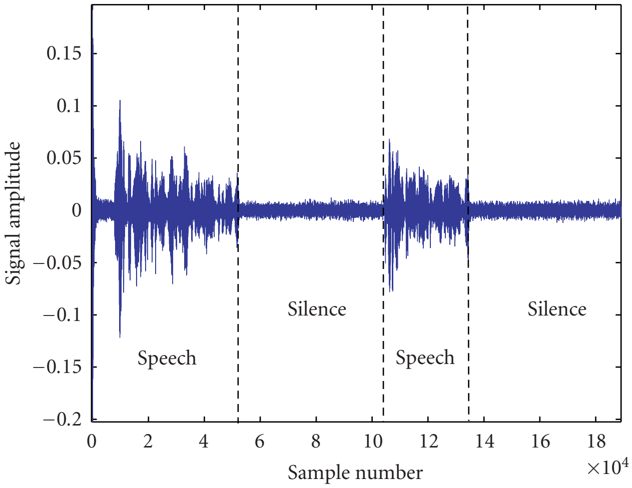

FIGURE 14: Audio waveform from a typical utterance in background noise. The speech and silence parts are hand-labeled to be used as ground truth.

7.1. Hardware

7.1.1. Audio sensors

The sensors consist of a microphone array. The audio signals are captured at 16 kHz on a Linux workstation using the advanced Linux sound architecture (ALSA) drivers. JACK is a useful audio server that is used here to capture and process multiple channels of audio data in real time (as required).

7.1.2. Video sensors

We use a synchronized pair of wide-angle cameras to capture the majority of the panorama around the table. The cameras are placed off the center of the table in order to increase their field of view as shown in the enlarged portion of Figure 9.

7.1.3. Synchronization

In order to facilitate synchronization, the video capture module generates a short audio pulse after capturing every frame. One of the channels in the microphone array is used to record this audio sequence and synchronize the audio and video frames.

7.2. Preliminary experiments and results

In order to evaluate the performance of the iterative decoding algorithm on a real-world problem, we consider a simplified version of the meeting room conversation, with one speaker. The goal of the experiment is to segment the speech data into speech and silence parts. The traditional approach to the problem is to use the energy in the speech signal as a feature and maintain an adaptive threshold for the energy of the background noise. This is not accurate in the presence of nonstationary background noise like overlapping speech from multiple speakers. In our experiment, we use the audio and video modalities to build a multimodal speech segmentation system, that is robust to background noise and which performs better than the video-only model or the joint model.

7.2.1. Data collection

We collected 4 minutes of audiovisual data from 20 different speakers. This included 12 different head poses and 2 different backgrounds as shown in Figure 10. We used 1 minute of data from each speaker, that is, a total of 20 minutes of audiovisual data, to estimate the HMM model parameters. The remaining 3 minutes from each speaker were included in the testing set. That is, a total of 60 minutes of testing data was used.

7.2.2. Feature extraction

Each time step corresponds to one frame of the video signal. The cameras capture video at 15 fps. We use the energy of the microphone signal in time window corresponding to each frame as the audio feature. We track the face of the speaker using the Viola-Jones face detector [20]. Figure 11 shows some sample frames from the face detector output for different subjects. We consider the mouth region as the lower half of the face. The motion in the mouth region is estimated by subtracting the mouth region pixels from consecutive frames and summing the absolute value of these differences. This sum is our video feature vector. Thus a smooth and stable face tracker is essential for accurate video feature extraction. Figure 12 shows the different positions of the lips during a typical utterance.

7.2.3. Modeling and testing

We then train HMMs in the audio and video domains using labeled speech and silence parts of speech data. We also

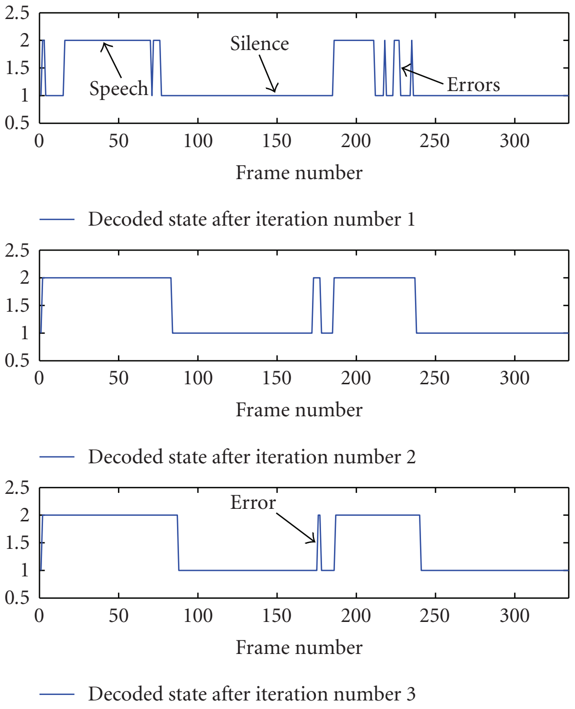

Figure 15: The decoded states of the HMM after each iteration. Note the errors in the first iteration being corrected in the subsequent iterations.

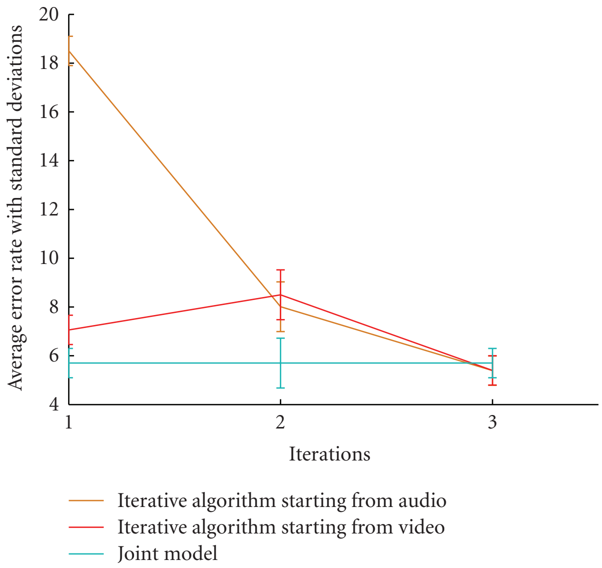

Figure 16: Results showing the error rates for the iterative decoding scheme for the speech segmentation problem.

construct the joint model by concatenating the features. The results of the experiment on a typical noisy segment of speech are shown in Figure 15. The ground truth is shown in Figure 14. From the numerical results in Figure 16, we see that by the third iteration, the iterative decoding algorithm performs slightly better than the joint model. This improvement, however, is not statistically significant because the background noise in the audio and video domains is not so severe. Though building the joint model is straightforward in this case, it is not so easy in more complex situations, as explained in the introductory sections. Thus the iterative algorithm appears to be a good fusion framework in the multimodal scenario.

8. CONCLUDING REMARKS

We have developed a general information fusion framework based on the principle of iterative decoding used in turbo codes. We have adapted the iterative decoding algorithm to the case of multimodal systems and demonstrated its performance on synthetic data as well as practical problem. In the future, we plan to further investigate its performance under different real-world scenarios and also apply the fusion algorithm to more complex systems.

We have also described the setup of an experimental testbed in the CVRR lab at UCSD. In the future, we plan to extend the experiments on the testbed to include many more features like speaker identification, affecting analysis and keyword spotting. This will lead to more complex human activity analysis tasks with more complex models. We will evaluate the effectiveness of the iterative decoding scheme on these complex real-world problems.

ACKNOWLEDGMENTS

The work described in this paper was funded by the RESCUE project at UCSD, NSF Award no. 0331690. We also thank UC Discovery Program Digital Media Grant for the assistance with the SHIVA lab testbed at Computer Vision and Robotics Research Laboratory (CVRR), UCSD. We acknowledge the assistance and cooperation of our colleagues from the CVRR Laboratory. We also thank the reviewers whose valuable comments helped us to improve the clarity of the paper.

REFERENCES

[1] M. M. Trivedi, K. S. Huang, and I. Mikić, “Dynamic context capture and distributed video arrays for intelligent spaces,” *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 35, no. 1, pp. 145–163, 2005.

[2] M. M. Trivedi, K. S. Huang, and I. Mikić, “Activity monitoring and summarization for an intelligent meeting room,” in *Proceedings of the IEEE International Workshop on Human Motion*, 2000.

[3] J. Ploetner and M. M. Trivedi, “A multimodal approach for dynamic event capture of vehicles and pedestrians,” in *Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks*, 2006.

[4] S. Oviatt, R. Coulston, and R. Lunsford, “When do we interact multimodally? Cognitive load and multimodal communication patterns,” in *Proceedings of the 6th International Conference on Multimodal Interfaces*, State College, Pa, USA, October 2004.

[5] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: a summary of the Johns Hopkins summer 2000 workshop," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, Cannes, France, 2001.

[6] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in *Proceedings of the 2nd International Conference on Audio-Visual Biometric Person Authentication*, Washington, DC, USA, March 1999.

[7] L. Chen, R. Malkin, and J. Yang, "Multimodal detection of human interaction events in a nursing home environment," in *Proceedings of the 6th International Conference on Multimodal Interfaces*, Pittsburgh, Pa, USA, October 2002.

[8] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, March 2005.

[9] T. Gustafsson, B. D. Rao, and M. M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[10] J. C. McCall and M. M. Trivedi, "Facial action coding using multiple visual cues and a hierarchy of particle filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2006.

[11] N. Nikolaidis, S. Siatras, and I. Pitas, "Visual speech detection using mouth region intensities," in *Proceedings of the European Signal Processing Conference*, Florence, Italy, September 2006.

[12] A. Jaimes and N. Sebe, "Multimodal human computer interaction: a survey," in *Proceedings of the IEEE International Workshop on Human Computer Interaction in Conjunction with ICCV*, Beijing, China, October 2005.

[13] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Journal of Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 270–287, 2006, special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

[14] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Proceedings of International Conference on Multimodal Interfaces*, Pittsburgh, Pa, USA, October 2002.

[15] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, "Integration of multimodal features for video scene classification based on Hmm," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, 1999.

[16] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.

[17] E. Erzin, Y. Yemez, A. M. Tekalp, A. Ercil, H. Erdogan, and H. Abut, "Multimodal person recognition for human-vehicle interaction," *IEEE Multimedia Magazine*, vol. 13, no. 2, pp. 18–31, 2006.

[18] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near shannon limit error-correcting coding and decoding: turbo-codes," in *Proceedings of the IEEE International Conference on Communications*, Geneva, Switzerland, May 1993.

[19] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, March 1974.

[20] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2002.