Triggering Memories of Conversations using Multimodal Classifiers

Wei-Hao Lin, Rong Jin, and Alexander Hauptmann

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213 USA
{whlin,rong,alex}@cs.cmu.edu

Abstract

Our personal conversation memory agent is a wearable 'experience collection' system, which unobtrusively records the wearer's conversation, recognizes the face of the dialog partner and remembers his/her voice. When the system sees the same person's face or hears the same voice it uses a summary of the last conversation with this person to remind the wearer. To correctly identify a person and help remember the earlier conversation, the system must be aware of the current situation, as analyzed from audio and video streams, and classify the situation by combining these modalities. Multimodal classifiers, however, are relatively unstable in the uncontrolled real word environments, and a simple linear interpolation of multiple classification judgments cannot effectively combine multimodal classifiers. We propose a meta-classification strategy using a Support Vector Machine as a new combination strategy. Experimental results show that combining face recognition and speaker identification by meta-classification is dramatically more effective than a linear combination. This meta-classification approach is general enough to be applied to any situation-aware application that needs to combine multiple classifiers.

Introduction

A memory agent is a wearable computer system that can provide information relevant to the current context of the wearer without user intervention. This context-sensitive information can serve as a passive reminder of things to do, or proactive association of past events. For example, Rhodes' remembrance agent [10] can provide text summaries based on current time, location, and conversation parter through a head-up display. Schiele et al.'s DyPERS [15] can retrieve video and audio that was previously associated with the physical object the wearer is watching. In this paper, we extend this work further by recognizing high-level objects such as specific faces and speakers, and remembering a summary of an earlier encounter.

Our research aims to develop a system that allows people to capture and retrieve from a complete record of their personal experiences. This assumes that within ten years technology will be in place for creating a continuously recorded, digital, high fidelity record of one's whole life in video form [4]. Wearable, personal digital memory systems units will record audio, video, location and electronic communications. This research aims to fulfill the vision of Vannevar Bush's personal Memex [1], capturing and remembering whatever is seen and heard, and quickly returning any item on request or based on the. While our vision outlines a research program expected to last for many years, we have reduced certain aspects of this vision into an operational personal memory prototype that remembers the faces and voices associated with a conversation and can retrieve snippets of that conversation when confronted with the same face and voice. The system currently combines face detection/recognition with speaker identification, audio recording and analysis. The face detection and speaker id enables the storing of the audio conversation associated with a face and a voice. Audio analysis and speech recognition compacts the conversation, retrieving only important phrases. All of this happens unobtrusively, somewhat like an intelligent assistant who whispers relevant personal background information to you when you meet someone you don't quite remember.

One key component in the aforementioned prototype is the combination of multimodal classifiers which determine the identity of the wearer's acquaintance as soon as the system recognizes a voice or a face. Multiple classifiers can improve the accuracy of the classification when the classifiers are complementary to each other. In tasks like person identification or validation, classifiers from different modalities use distinct features to classify samples, and thus they seldom make correlated mistakes at the same time. Using an ensemble of multimedia classifiers has been previously researched in identity verification studies [3][7], which demonstrated the effectiveness of linearly combining up to three multimodal classifiers. In order to achieve a high accuracy of classification, the strategy of combining evidence from a group of classifiers plays a critical role. Majority voting and linear interpolation [5] are the most common ways of combining classifier output. However, these methods only utilize the final decision from each classifier, ignoring the complete picture of all judgments. The other problem is that the weights between classifiers are assigned equally (summing all probabilities) or fixed empirically based on the reliability of each classifier. To address these problems, we propose a new

combination method called meta-classification, which makes the final decision by re-classifying the result each classifier returns. Experiment results show meta-classification is more effective than weighted linear combinations.

This paper is organized as follows: Our system for collecting and the retrieving digital human memory is described in Section 2. The multimedia classifiers are detailed in Section 3. Section 4 explains meta-classification, which is our proposed new combination strategy. Experimental results are given in Section 5 and conclusions presented Section 6.

Personal Conversation Memory Agent

There are currently two modes of system operation: Memory collection (learning) and memory retrieval.

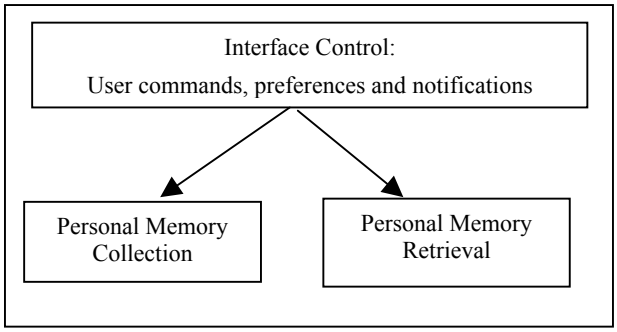

Figure 1. The basic architecture of the Personal Memory system

The basic hardware components of the system are a wearable miniature digital video camera, 2 microphones (1 close-talking and 1 omni-directional), headphones to receive user output and a laptop type computer for processing. The software modules involved in the system are a module for speaker identification, a speech recognition module, a face detection and recognition module, a database, and an interface control manager module. The basic architecture of the system, as outlined in Figure 1, shows the interface control manager selecting between the acquisition and the retrieval module as needed.

Interface Control

The interface control module determines which of 3 states the system is currently in: Idle, Collecting Memory or Retrieving Conversations from Personal Memory. The current user input interface to this control module consists of a wearable mouse, with one button to start collecting conversation memory, one button to start retrieving the conversations (or to skip to the next relevant one) and one button to set the system into an idle state where neither collection nor retrieval takes place.

On the output side, it is assumed that all output will be provided through the headphones to the wearer. A visual display interface is available mostly for test and demonstration purposes. The interface control has the ability to provide a set of short audio notifications to inform the user what the system is currently doing. These notifications include:

- Memory collection is starting
- Memory collection is ending
- Memory retrieval is starting
- Memory retrieval is ending
- Face was detected in audio stream
- Failure to detect faces
- Various internal system failure states

Our experience has been that these user notifications were extremely important to keep the user informed of what the system is doing, since there is no visual feedback in normal use.

Future versions of the system will likely have the interface manager module controlled through a limited vocabulary, command speech interface (e.g. “Record memory” or “Who is this person?”). We envision that eventually the interface system will be triggered entirely by context-dependent recognized dialogue phrases, for example, ``Nice to meet you", “My name is”, etc., instead of mouse clicks or specific commands by the user.

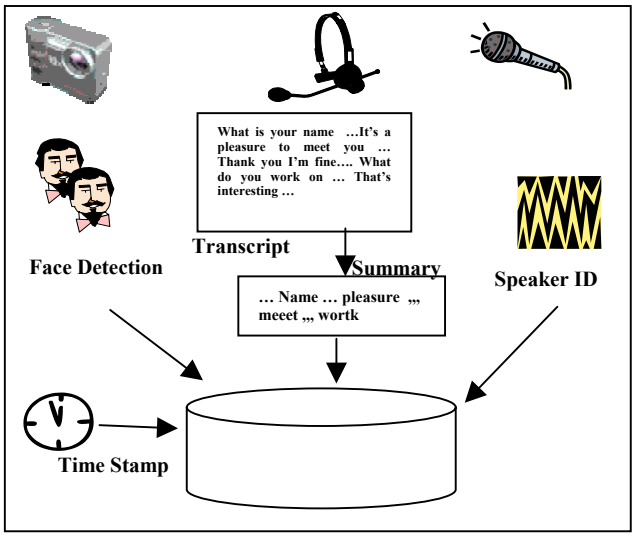

Figure 2. Process for Personal Memory Collection

Personal Memory Collection

The system works as a wearable device consisting of a miniature ‘spy’ camera, a cardioid lapel microphone and an omni-directional microphone all attached to a laptop computer. The system works by detecting the face of the person you are talking to in the video, and listening to the conversation from both the close-talking (wearer) audio track and the omni-directional (dialogue partner) audio track. An overview of the ‘learning’ system for memory collection is shown in Figure 2.

The close-talking audio is transcribed by a speech recognition system to produce a rough, approximate transcript. The omni-directional audio stream is processed through a speaker identification module. An encoded representation of the face of your current dialog partner, the dialog partner speaker characteristics, and the raw audio of the current conversation is saved to a database. The next time the system sees the same person (by detecting a face and matching it to the stored faces in the database), it can retrieve and play back the audio from the last conversation.

The audio can optionally be processed through audio analysis (silence removal, emphasis detection) and general speech recognition to efficiently replay only the person names and the major issues that were mentioned in the conversation.

Personal Memory Retrieval

In the retrieval (remembering) mode, the system immediately searches for a face in the video stream and performs speaker identification on the omni-directional audio stream. Once a face is detected, the face and speaker characteristics will be matched to the instances of faces and speaker characteristics stored in the memory database. The score of both faced and speaker matches is combined using our meta-classification strategy. When a sufficiently high scoring match is found, the system will return a brief summary of the last conversation with the person. Figure 3 shows the process of personal memory retrieval.

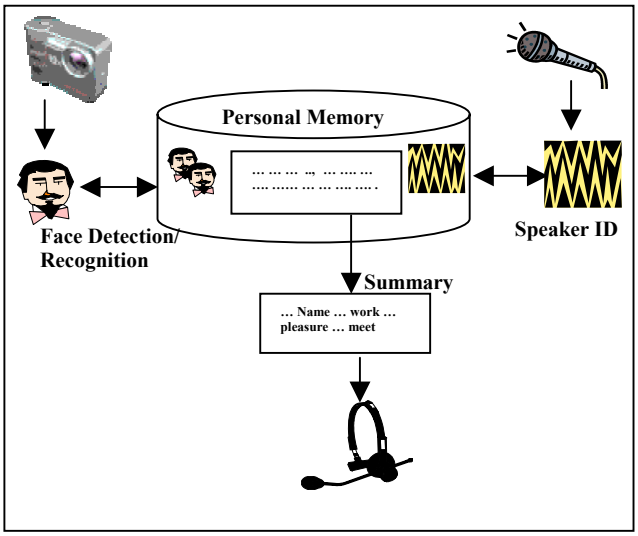

Figure 3. Process for Personal Memory Retrieval

Multimodal Classifiers

Face Detection and Recognition

Extensive work in face detection has been done at CMU by Rowley [11][12][13]. This approach modeled the statistics of appearance implicitly using an artificial neural network. Currently we use Schneiderman's approach [18], which applies statistical modeling to capture the variation in facial appearance. We learn the statistics of both object appearance and "non-object" appearance using a product of histograms. Each histogram represents the joint statistics of a subset of wavelet coefficients and their position on the object. Our approach is to use many such histograms representing a wide variety of visual attributes. The detector then applies a set of models that each describes the statistical behavior of a group of wavelet coefficients.

Face matching was used in [14] with the 'eigenface' approach. Meanwhile there have been several commercial systems offering face detection and identification, such as Visionics [19]. In our implementation we have been using both the Visionics FaceIt toolkit for face detection and matching as well as the Schneiderman face detector and 'eigenfaces' [18] for matching similar faces. Eigenfaces treat a face image as a two-dimensional N by N array of intensity values. From a set of training images, a set of eigenvectors can be derived that constitute the eigenfaces. Every unknown new face is mapped into this eigenvector subspace and we calculate the distance between faces through corresponding points within the subspace [20].

Speaker Identification

Speaker identification is done through an implementation of Gaussian Mixture Models (GMM) as described by Gish [16]. Gaussian Mixture Models have proven effective in speaker identification tasks in large databases of over 2000 speakers [7][16].

Prior to classification, Mel-Frequency Cepstral Coefficients (or MFCC) features are extracted from the audio channel. For training, regions of audio are labeled with a speaker code, and then modeled in their respective class (speaker). Once training models have been generated, the system must classify novel audio sections. The process begins by segmenting the audio channel into 1-second, overlapping regions and computing the GMM. The resulting model is compared to existing trained models using a maximum likelihood distance function. Based on the comparisons to each class, a decision is made as to the classification of the data into speech, noise, known speaker X, etc. The speaker identification system also uses the fundamental pitch frequency to eliminate false alarms. Generally, about 4 seconds of speech are required to get reliable speaker identification, under benign environmental conditions.

Combining Classifiers

The main idea of meta-classification is to represent the judgment of each classifier for each class as a feature vector, and then to re-classify again in the new feature space. The final decision is made by the meta-classifiers instead of just linearly combining each classifier's judgment. In this section, we first introduce the generation of the new features, followed by a description of the widely used linear interpolation method, then finally our method of building a meta-classifier.

Feature Synthesis

Multimedia classifiers make judgments at different time periods because of discrepant characteristics of the individual modalities. For example, the speaker identification module usually takes longer to report a result than the face recognition module because the former works with a larger time window while the latter can make a judgment as soon as an image is ready to be analyzed. Consequently, the classification results from these multimodal classifiers will be fed into the meta-classifier asynchronously, and a method of combining them appropriately is needed.

Given an example with an unknown class, set x_j^i is the degree of likelihood that the example belongs to the class i as made by the classifier j. Depending on the nature of the classifier, x_j^i can be a similarity score or probability. A multimedia classifier j generates a classification vector $\boldsymbol{x}_j$ once it finishes analyzing input from the audio or video stream. The classification vector $\boldsymbol{x}_j = (x_j^1, x_j^2, ..., x_j^k)$, where k is the number of classes, i.e. the number of people in current pool of digital human memory. Given r classifiers, at a time point t at which any classifier can make a judgment, a new feature vector $\boldsymbol{x}_{syn}(t) = (\boldsymbol{x}_1(t'), \boldsymbol{x}_2(t'), ..., \boldsymbol{x}_r(t'))$, where t' is the time point that is the closest to time point t when the classifier makes a judgment. In other words, whenever a classifier makes a judgment, a new feature vector combines every classifier's judgment by concatenating classification vectors generated at the time point nearest to current time. Suppose we have three classifiers and two classes. At the fifth second, the first classifier makes a classification judgment $x_1(5) = (10, 2)$. The most recent judgments made by the other two classifiers are $x_2(4.375) = (0.8, 0.1)$ and $x_3(3) = (60, 50)$ at 4.375 and 3 seconds, respectively. Therefore, the synthesized vector is $x_{syn}(5) = (10, 2, 0.8, 0.1, 60, 50)$, and the meta-classifier learns in this new feature space. The synthesis method is based on the assumption that between t and t', there is no dramatic change with respect to the last judgment, which holds true in the current situation. When the user attempts to retrieve an acquaintance from memory by matching the face or voice, he or she usually continues to look or listen to the person who is to be identified.

Linear Combination

Kittler et al. [8] proposed a probability framework to explain various schemes of combining judgments from multiple classifiers. If each person in our conversation memory system represents a class $(w_1, w_2, ..., w_p)$, where p is the number of people in the memory pool, the task of identifying the unknown person in the video or audio stream can be formulated as classifying the person by combing multimodal features x_i, $i = 1, ..., r$, where r is the number of multimodal classifiers. Based on the assumption that each classifier is conditionally independent, the decision rule classifies the unknown person into the class w_j if

$$P^{-(r-1)}(w_j)\prod_{i=1}^{r} P(w_j \mid x_i) = \max_{k=1}^{p} P^{-(R-1)}(w_k)\prod_{i=1}^{r} P(w_k \mid x_i)$$

Furthermore, under the assumption that the posterior probability $p(w|x)$ is not very far from the prior probability $p(w)$, the decision rule becomes a sum of posterior probabilities as follows,

$$(1-r)P(w_j) + \sum_{i=1}^{r} P(w_j \mid x_i) = \max_{k=1}^{r}\left((1-r)P(w_k) + \sum_{i=1}^{r} P(w_k \mid x_i) \right)$$

The sum rule combines classifiers by summing the judgments made by each classifier for each class, and the class with the highest probability is chosen as the final decision. It may be counter-intuitive at first to assume the posterior probability and prior probability are close, but the sum rule is widely used as a combination scheme, and outperformed other linear combination strategies such as max rule, min rule, median rule, and majority voting [8]. In this paper, we use the sum rule with equal prior probabilities for comparison against our new method.

One constraint imposed by the sum rule, or other probability-based combination strategies, is that each classifier must express its decision as a true probability. Since not every classifier is designed to output probability, we transform their judgments into probability by normalizing the total sum of the similarity scores generated by the classifier, i.e.

$$p(w_i \mid x_j) = \frac{f(x_j, w_i)}{\sum_{k=1}^{p} f(x_j, w_k)}$$

where $f(x, w)$ is a function of measuring the similarity between x and the class w made judged by each classifier.

Meta-Classification

Meta-classification is re-classifying the classification results made by classifiers. Consider that there is one 'deaf' face recognition experts and two 'blind' speaker identification experts residing in our system. Once the system detects an unknown person approaching the user or the user actively triggers the recognition mode, each expert starts to make his or her own decision based on the input from the corresponding modality. Instead of making the

final decision by voting, or summing up probabilities and then picking the most promising one, we present their decisions as a synthesized vector to another judge, i.e. the meta-classifier, who ultimately decides the identity of the person in the current video or audio stream.

A very promising classification technique, Support Vector Machine (SVM) [2] is used here as a meta-classifier. The basic of idea of SVM is to separate samples with a hyperplane that has a maximal margin between two classes. To formulate the problem of classifying synthesized feature vectors, the training data are represented as $\{\boldsymbol{x}_i, y_i\}$, $i = 1, 2, \ldots, R$, y_i is either -1 (negative examples) or 1 (positive examples), R is the number of training samples. Suppose all training data satisfy following constraints:

$$\begin{aligned} \boldsymbol{x}_i \bullet \boldsymbol{w} + b \geq +1 \text{ when } \boldsymbol{y_i} = 1 \\ \boldsymbol{x}_i \bullet \boldsymbol{w} + b \leq -1 \text{ when } \boldsymbol{y_i} = \text{-}1 \end{aligned} \quad (1)$$

The distance between the hyperplane $\boldsymbol{x}_i \bullet \boldsymbol{w} + b = 1$ and the hyperplane $\boldsymbol{x}_i \bullet \boldsymbol{w} + b = -1$ are $2/\|\boldsymbol{w}\|$, where $\|\boldsymbol{w}\|$ is the Euclidean norm of $\boldsymbol{w}$. Therefore, by minimizing $\|\boldsymbol{w}\|^2$ we get the two hyperplanes with maximal margins. Quadratic programming provides well-studied optimizations to maximize the quadratic functions subject to the linear constraints in Equation 1, which guarantees finding the global maximum. SVM is not only theoretically sound, but outperforms other classification algorithms in empirical problems with high dimensionality [17]. The SVM meta-classifier makes its binary decision by classifying synthesized feature vectors, and we build one such meta-classifier for each class. The maximal margin suggests having better generalization ability. Unlike other combination schemes that require each classifier to have the same output form, here feature vectors can consist of scores or similarities without any restriction. We implement the SVM meta-classifier using SVM$^{\text{light}}$ [6] with a linear kernel.

The advantage of applying meta-classification is two-fold. First, when combining multiple classifiers, the similarity score or probability produced by each classifier does not necessarily convey all of the information. The distribution of the scores for each class judged by the classifier reveals how confident it is in making the decision, which is a characteristic that can only be captured by a classification feature vector $\boldsymbol{x}_r$ but not in normal combination schemes such as linear interpolation. Second, there may be some patterns across several classification vectors, which can be learned by a meta-classifier. For example, one of the users' friends was first met in a very noisy environment, resulting in poor quality voice for training speaker identification but keeping the visual features of face intact. Meta-classification can learn the pattern from synthesized feature vectors. Therefore, when the user meets the friend again, the face recognition module will be certain about identifying the friend while the voice recognition module is confused. The normal linear combination strategy will act unstable in this circumstance. The meta-classifier, on the contrary, can make a better decision by observing the patterns in the results from the multimodal classifiers.

The meta-classification strategy can be applied to other classifiers with little effort. Any existing multimedia classifier can be plugged into the framework to combine with other classifiers to generate synthesized feature vectors, and meta-classification training is processed in the same way. It does not matter that the classifier is probability-based or similarity-based, and both probabilities and similarity scores can be combined into the feature vector.

Evaluation Window

To exploit the continuousness of audio and video input in a context-aware application, we can make the classification decision not only by combining multimodal classifiers, but also accumulating classification results over time. The number of times of classification judgments is defined as the size of the evaluation window. The decision rule that we choose subject j as the final decision is as follows:

$$\sum_{t=i}^{i+w} s_j(t) = \max_k \sum_{t=i}^{i+w} s_k(t)$$

where $s_k(t)$ is the judgment (probability or similarity scores) the classifier made for the k-th class at the time t, i is the starting time when we begin to accumulate judgments, and w is the size of the window.

Experiment

Data Collection and Procedure

We collected two conversations each with 22 people while wearing our prototype memory capture unit. Each conversation was at least 20 seconds long, and was analyzed for faces and speaker audio characteristics as described above. The lighting condition and background were different between the two conversations. The first of each conversation served as the training example for multimedia classifier, while the second conversation was used as a query or retrieval prompt to 'remember' the first conversation. The retrieval was considered successful only when the combining strategy correctly identified the person in question.

The first conversation was used to train the face detection/recognition and speaker identification classifiers, and the meta-classifiers as well. The second conversation was used to test the performance. Since the SVM meta-classifier is a binary classifier, we have to train a meta-classifier for each of the 22 people. The training data consisted of synthesized feature vectors from the given person, and feature vectors from the other 21 people. To account for the discrepancy between the number of positive examples and negative examples, the cost of misclassifying positive training examples into negative examples was 22 times the cost incurred in the reverse situations. There were 9316 testing feature vectors for the total 22 classes. Note that 9316 is not multiple of 22 because the number of feature vectors generated from each person was not the

same. If one of the multimedia classifier had hard time making judgment at a time slot, there would not be a feature vector synthesized at that time point.

Result

We used the average rank as the evaluation metric, i.e. on average, at what rank was the correct conversation found. The better the classifier or the combining strategy performs, the closer its average rank is to one. The results of our experiment is shown in Table 1, suggesting that the Visionics face recognition system found the correct conversation at rank 3.33 of the 22 possible conversation candidates. Speaker identification by acoustic MFCC similarity proved to be more reliable with an average rank of 3.92, and the accuracy of speaker identification by pitch was the worst at 6.22 among all the multimodal classifiers. The meta-classifier combined the face classifier and the speaker identification, and resulted in an average rank of 2.61. All the above results are calculated with the size of the evaluation window of one, which means no information over time is used. The sum rule does not outperform single classification as previously suggested, and achieved a performance between face recognition and speaker identification. The result of meta-classifier does not only significantly outperform individual multimedia classifiers, but also outperforms the sum rule combination strategy.

Classifiers	Average Rank
Visionics Face Recognition	3.33
Speaker ID with Similarity	3.92
Speaker ID with Pitch	6.22
Combine using Sum rule	3.87
Combine using SVM meta-classification	2.61

Table 1 Experimental results from each classifier and combination strategies (window size = 1) showing the advantage of meta-classification

We also evaluated the effect of window size, and the plot of the classifiers and combination strategies versus the window size is shown in Figure 4. We expect that with the increasing size of the evaluation window, the performance, i.e. average rank, should improve because the classifier is more confident about its decision by observing more situation over time. Since classifiers from different modalities make decisions at different pace, the plot in Figure 4 has two x-axes, the one above the plot with fewer window sizes for 25 seconds is for slow data rate speaker identification, and the bottom one with more window sizes is for the faster data rate of face recognition and the combination strategies. Interestingly, the two audio classifiers do not improve with the size of window, which suggests that the speaker identification modules have stable, but not very accurate performance with a one-second audio sample. The performance may not improve unless the sample size expands, which is not tolerable in context-aware applications that need a quick response from each modality. On the other hand, the face recognition module and the combination strategies improve with the size of the evaluation window. Note that after a window size of 250 (corresponding to about 15 seconds of video), the meta-classifier achieved the perfect performance. Moreover, the meta-classifier combining strategy showed the curve declining quickly in the first several window sizes, which means the strategy is effective at combining multimodal classifiers to make the best classification judgment to retrieve the correct conversation. To achieve an average rank of two, the meta-classification strategies only need a window size of 20 (about 5 seconds).

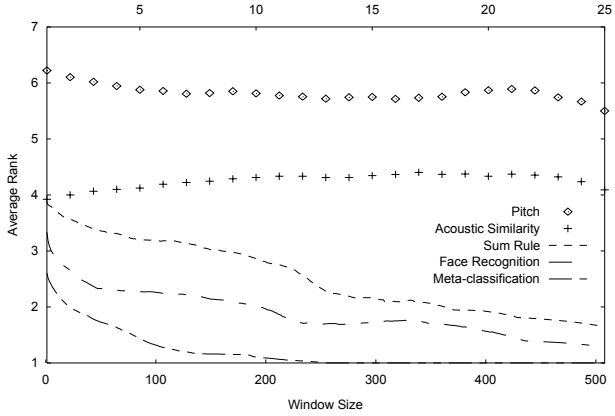

Figure 4 Experimental results of manipulating the size of the evaluation window, which shows the rapid convergence of the SVM meta-classification

Conclusion

We present a wearable conversation collection system that unobtrusively records a conversation in audio and video, builds an association between the partner's biometric (audio/video) characteristics and that conversation, and automatically retrieves the previous conversation once the person is seen by the camera or is heard via the microphone. This prototype is based on the critical assumption that the system can effectively and efficiently identify an unknown person by combining evidence from multiple modalities, including face recognition and speaker identification. We proposed a novel meta-classification strategy of combining multimedia classifiers. Based on the experiment results in this task of identifying the same person through audio and video signals, meta-classification was shown to be much more effective than single classifiers as well as a linear combination strategy. The result also showed that meta-classification strategy could improve quickly with the increase size of evaluation window. Although the result is not perfect in terms of speed and classification result, we expect the results can be improved, as more context information is included. With the emerging need of combining different modalities in situation-aware applications, our method can provide a general framework

to integrate different sources of context information, and provide more confident classification judgments.

Reference

[1] V. Bush, "As we may think", *Atlantic Monthly*, Vol.176, No. 1, pp. 101-108, 1945

[2] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", *Knowledge Discovery and Data Mining*, Vol. 2, No. 2, 1998

[3] R. W. Frischholz and U. Dieckmann, "BioID: A Multimodal Biometric Identification System", *IEEE Computer*, Feb 200, pp. 64-68

[4] J. Gray, "What next? A few remaining problems in Information Technology", *ACM Federated Research Computer Conference*, Atlanta, GA, May 1999

[5] S. Hashem and B. Schmeiser, "Improving Model Accuracy using Optimal Linear Combinations of Trained Neural Networks", *IEEE Transactions on Neural Networks*, Vol. 6, No. 3, May 1995, pp. 792-794

[6] T. Joachims, "Making large-Scale SVM Learning Practical" , *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999

[7] O. Kimball, M. Schmidt, H. Gish, andJ. Waterman, "Speaker verification with limited enrollment data", *ICCSLP-96, International Conference on Spoken Language Processing*, volume 2, pages 967 – 970, Philadelphia, PA, 1996.

[8] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas, "On Combining Classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, Mar 1998

[9] A. Pentland , and T. Choudhury "Face Recognition for Smart Environments", *IEEE Computer*, Feb 2000, pp. 50-55

[10] Bradley J. Rhodes, "The Wearable Remembrance Agent: a System for Augemented Memory", *Perosnal Technologies*, 1(1)

[11] H. A. Rowley, S. Baluja, and T. Kanade "Neural Network-Based Face Detection", *IEEE Transaction on Pateern Analysis and Machine Intelligence*, Jan 1998, pp. 23-38

[12] H. A. Rowley, S. Baluja and T. Kanade, "Human Face Detection in Visual Scenes", Carnegie Mellon University, *Technical Report CMU-CS-95-158*, Pittsburgh, PA

[13] H. A. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection", *IEEE CVPR*, Santa Barbara, 1998.

[14] M. Turk and A. Pentland, "Eigenfaces for recognition", *Journal of Cognitive Neuroscience*, 3(1):71—86, 1991.

[15] B. Schiele, N. Oliver, T. Jebara, and A. Pentland, "An interactive computer vision system, DyPERS:dynamic and personal enhanced reality system", *International Conference on Computer Vision Systems*, 1999

[16] M. Schmidt, J. Golden, and H. Gish "GMM sample statistic log-likelihoods for text-independent speaker recognition," *Eurospeech-9*, Rhodes, Greece, Sep 1997, pp. 855-858

[17] B. Scholköpf, K.-K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers", *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, Nov 1997

[18] H. Schneiderman and T. Kanade, "Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition", *IEEE CVPR*, Santa Barbara, 1998

[19] Visionics FaceIt Developer Kit, http://www.visionics.com

[20] S. Satoh and T. Kanade, Name-It: Association of Face and Name Video, tech. report CMU-CS-96-205, Computer Science Department, Carnegie Mellon University, 1996.