Meta-classification of Multimedia Classifiers

Wei-Hao Lin
*Carnegie Mellon University*

Rong Jin
*Carnegie Mellon University*

Alexander Hauptmann
*Carnegie Mellon University*

Follow this and additional works at: [http://repository.cmu.edu/compsci](http://repository.cmu.edu/compsci)
Meta-classification of Multimedia Classifiers
Wei-Hao Lin, Rong Jin, Alexander Hauptmann
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{whlin,rong,alex}@cs.cmu.edu

ABSTRACT
Combining multiple classifiers is of particular interest in the multimedia systems, since there is usually data of very different types/modalities that should be mined or analyzed. Our wearable ‘experience collection’ system 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 can then use a summary of the last conversation with this person to remind the wearer. To correctly identify a person from a mixture of video and audio stream, classification judgments from individual modality classifiers must be combined effectively to yield a more accurate decision. A meta-classification strategy of combining multimodal classifiers using Support Vector Machine is proposed. Preliminary, empirical results show that combining different face recognition and speaker identification technology by meta-classification is dramatically more effective than weighted interpolation. Meta-classification is general enough to be applied to any application that needs to combine multiple classifiers without much modification.

Keywords
Meta-classification, Classifier Combination, Multimedia Classification, Face Recognition, Speaker Identification

1. INTRODUCTION
Classification is an important form of knowledge extraction, and can help make key decisions. Multimedia classification works on visual or aural features that are quite different from text and numerical data, and needs to integrate results from algorithms in fields such as computer vision, pattern recognition, speech recognition, and machine learning. Visual classifiers for fingerprints [6], face recognition [11], [12], iris matching [9], and audio classifiers for speaker identification [15] have been successfully applied to many domains like security surveillance, optical character recognition, and smart environments [10].

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 [1]. 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. While our vision outlines a research program expected to last for few 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 such that the identity of the wearer’s acquaintance can be determined 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 individual multimodal classifiers has been previously researched in identify verification studies [7][2], which demonstrated the effectiveness of linearly combining up to three multimodal classifiers. However, to achieve high accuracy of classification, the strategy of combining evidence from a group of classifiers plays a vital role. Majority voting and linear interpolation [5] are the most common way of combining classifier output. In this paper we propose a new combination method called meta-classification, which makes the final decision by re-classifying the result each classifier returned. Experimental results show meta-
classification is more effective than weighted linear combinations.

This paper is organized as follows: Our basic 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.

OUR ‘PERSONAL MEMORY SYSTEM’

We have implemented an initial prototype of a personal memory system that functions like the previously described intelligent assistant.

There are currently two modes of system operation: personal conversation data collection and remembering.

1.1 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.

1.2 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.

2. MULTIMEDIA CLASSIFIERS

2.1 Face 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 [16], 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 describe the statistical behavior of a group of wavelet coefficients.

Face matching was used in [13] with the ‘eigenface’ approach. Meanwhile there have been several commercial systems offering face detection and identification, such as Visionics [18]. In our implementation we have been using both the Visionics Facet toolkit for face detection and matching as well as the Schneiderman face detector and ‘eigenfaces’ [16] 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 [19].

2.2 Speaker Identification

Speaker identification is done through our own implementation of Gaussian Mixture Models (GMM) as described by Gish [15]. Gaussian Mixture Models have proven effective in speaker identification tasks in large databases of over 2000 speakers [20, 21].

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.

3. 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 introduce the generation of the new features first, followed by the method of building a meta-classifier.

3.1 Feature Synthesis

Multimedia classifiers make judgment at different time periods because of discrepant characteristics of modalities. For example, the speaker identification module usually takes a longer time to report than the face recognition module because the former works in the time domain while the latter can make a judgment as soon as an image is ready to be analyzed. Consequently, the classification results from 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 \( x^j \) 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, \( x^j \) can be a similarity score or probability. A multimedia classifier \( j \) generates a classification vector \( x_j \) once it finishes analyzing input from the audio or video stream. The classification vector \( x_j = (x_j^1, \ldots, x_j^k) \), 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 \( x_{syn} \) can be synthesized by Equation 1.

\[
X_{syn}(t) = (X_1(t'), X_2(t'), \ldots, X_r(t'))
\]

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 meta-classifier can learn using 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 context. When
the user attempts to retrieve a previous acquaintance from memory by matching face or voice, he or she usually continues to look or listen to the person who is to be identified.

3.2 Meta-classification

Meta-classification is re-classifying the classification results made by classifiers. Consider that there are two ‘deaf’ face recognition experts and one ‘blind’ speaker identification expert 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 idea of SVM is to separate samples with a hyperplane that has a maximal margin between two classes. SVM is based on statistical learning theory, and, for quadratic programming problems, the training is guaranteed to find the global maximum. SVM is not only theoretically sound, but outperforms other classification algorithms in empirical problems with high dimensionality [16]. The SVM meta-classifier makes its binary decision by classifying synthesized feature vectors, and we build one such meta-classifier for each class. 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. SVMlight [7] was adopted as the implementation of meta-classifiers.

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 $\mathbf{x}$, but not in normal combining 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 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 pattern in the results from multimodal classifiers.

4. EXPERIMENT

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 two conversations. The first of each conversation served as the training example for multimedia classifier, while the second conversation was used as query or retrieval prompt to ‘remember’ the first conversation. The retrieval was considered successful only when the combining strategy makes correct classification of the person in question.

The second conversation was divided into two parts: the first 15 seconds were used to generate synthesized feature vectors and to assign the correct labels to train the meta-classifiers, and the last 5 seconds were used to test each multimedia classifier as well as to generate synthesized feature vectors to test the meta-classifier. There were 345 testing feature vectors for the 22 classes. Note that 345 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. On the average, the face classifier made a judgment every 1/6 second, and the voice classifier made a judgment every second.

4.1 Result

We used the average rank as the retrieval 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 Figure 2, suggesting that the Schneiderman/Eigenface detection/recognition method retrieved the correct conversation at an average rank of 3.42 of the 22 possible conversation candidates. The Visionics face recognition system found the correct conversation only at rank 4.5. Speaker identification by acoustic MFCC similarity proved to be slightly more reliable with an average rank of 3.09, but the accuracy of speaker identification by pitch was between that of the two face recognition classifiers. The meta-classifier combined the face classifiers and the speaker identification. The result of meta-classifier does not only significantly outperform individual multimedia classifiers, but also outperforms other strategies of linearly combining or summing classifiers.
5. CONCLUSIONS

Since this is a preliminary study, we are fully aware of problems that exist with our current data and test procedure. While no test data was used for training, since parts of the 2nd conversation were used for the SVM training, certain environmental factors would have been similar to the test data not present in the other classifiers. However, we found that a two-classifier SVM trained using only the 1st conversation still improved accuracy over the best linear combination. Therefore, we believe these results will hold up in more rigorous, larger scale evaluations and look forward to presenting them at the workshop.

The novelty of our research is the meta-classification strategy of combining multimedia classifiers. Based on the experimental results in the task of identifying the same person through audio and video signals, meta-classification is shown to be much more effective than single classifiers as well as any linear combination strategy and summing-up strategy. The meta-classification combination strategy can be easily applied to other situations that need to combine multiple classifiers to improve classification accuracy.

Why is SVM better. (repeat motivation ideas)

6. REFERENCES


Table 1 Experimental Results from individual classifiers and the combinations of classifiers showing the advantage of SVM.