Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2005-081
AIM-2005-35
CBCL-258

December 14, 2005

Error Weighted Classifier Combination for Multi-modal Human Identification

Yuri Ivanov, Thomas Serre, Jacob Bouvrie

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Error Weighted Classifier Combination for Multi-modal Human Identification

Yuri Ivanov
Honda Research Institute, US
145 Tremont St
Boston, MA 02111

Thomas Serre
MIT CBCL
45 Carleton St
Cambridge, MA 02139

Jacob Bouvrie
MIT CBCL
45 Carleton St
Cambridge, MA 02139

Abstract

In this paper we describe a technique of classifier combination used in a human identification system. The system integrates all available features from multi-modal sources within a Bayesian framework. The framework allows representing a class of popular classifier combination rules and methods within a single formalism. It relies on a "per-class" measure of confidence derived from performance of each classifier on training data that is shown to improve performance on a synthetic data set. The method is especially relevant in autonomous surveillance setting where varying time scales and missing features are a common occurrence. We show an application of this technique to the real-world surveillance database of video and audio recordings of people collected over several weeks in the office setting.

1 Introduction and Motivation

In problems of biometric verification and identification a large role is played by the multi-modal aspect of the observation. A person can be identified by a number of features, including face, height, body shape, gait, voice etc. However, the features are not equal in their overall contribution to identifying a person. For instance, modern algorithms for face classification (e.g. [11]) and speaker identification (e.g. [6]) can attain high recognition rates, provided that the data is well formed and is relatively free of variations and noise, while other features, such as, gait (e.g. [1]) or body shape, are only mildly discriminative.

Even though one can achieve high recognition rates when classifying some of these features, in reality they are observed only relatively rarely - in a surveillance video sequence the face image can only be used if the person is close enough and is facing the camera, or a person's voice when the person is speaking. In contrast, there is a plentiful supply of the less discriminative features. This situation is illustrated on an example of one of our video sequences in figure 1.

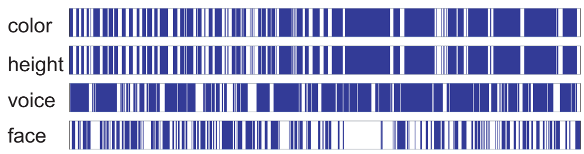

Figure 1: Illustration of feature availability for a long video sequence. Presence of the feature in the corresponding input frame is indicated by color.

A classifier derived from weakly discriminative features is usually highly inaccurate. This inaccuracy is determined by parameters external to the classifier, such as noisy measurements, or by an intrinsic inadequacy of the chosen method. If, for instance, we were to choose an individual's height as a discriminating measure, several people could have approximately the same height, and therefore, look inherently alike to the height classifier. Luckily the intrinsic inadequacy can be measured from the training data and used to subsequently assign a confidence measure which weights the classifier output in combination.

In pattern recognition and voice verification there has been significant interest in using multiple classifiers in order to improve the recognition rate of a given classification system. Many comparisons have been made between alternative combination rules, such as sum and product rules. In particular, [9, 7, 4, 5]. Bilmes and Kirchhoff, [2], as well as Tax *et. al.*, [10] point out that the product rule is optimal when the classifiers in the ensemble are correlated, while sum (or mean) rule is preferred if they are not. Rank order statistics rules (e.g. min/max) are more robust to outliers than the sum rule, but typically do not offer as much improvement over the error variance [5]. Voting schemes are often more robust to non-Gaussian or multimodal distributions of the data.

In this paper we present a general framework for combining classifiers in which many schemes of classifier combination may be represented. We examine properties of this framework on synthetic data and then apply it in the context of our automated on-line human identification system.

2 Bayesian view of classifier combination

Typically, a fully trained classifier misclassifies at least some of the training data. These misclassifications are reflected in the form of a *confusion matrix*. The confusion matrix expresses the likeness of the classes from the point of view of this particular classifier, which it is trained to recognize. This matrix represents an empirical value of the distribution of the intrinsic error of the classifier on the given data set. Our approach to the classifier combination is based on using the conditional error distribution derived from the confusion matrix to weigh the output of each classifier before the application of a combination rule.

In more rigorous terms, a set of features x in our setting, represents measurements available from multiple independent observation channels. That is $x = \left\{x^{\lambda=1}, x^{\lambda=2}, ...x^{\lambda=C}\right\}$, where C is the number of individual feature channels, such as an image of the person's face, person's height, distribution of colors in the person's clothes, etc. Our goal is for a given observation set, x, to infer a true class label, ω, which takes values from 1 to K, the number of classes.

Since each classifier in the set uses only a disjoint subset x^λ of the features in x, we can assert that $\forall\lambda : P(\omega|x,\lambda) \equiv P(\omega|x^\lambda,\lambda)$. Then for a full observation, x, the output of the classifier system, ω, can be expressed in terms of a marginal distribution:

$$P(\omega|x) = \sum_{i=1}^{C} P(\omega, \lambda_i|x) = \sum_{i=1}^{C} P(\omega|\lambda_i, x)P(\lambda_i|x) \quad (1)$$

where $P(\lambda_i|x)$ is the weight assigned to i-th classifier in the combination. In different formulations this term represents an "expert" or a "critic".

Our framework for classifier combination is based on viewing the output of an individual classifier as a random variable, $\tilde{\omega}$. Suppose that for each classifier λ_i we have access to the joint probability of the true and predicted class labels, $P(\omega, \tilde{\omega}|x, \lambda_i)$. Then the true label can be inferred from the individual classifier by averaging with respect to the classifier prediction:

$$\begin{aligned} P(\omega|\lambda_i, x) &= \sum_{k=1}^{K} P(\omega, \tilde{\omega}_k|\lambda_i, x) \\ &= \sum_{k=1}^{K} P(\omega|\tilde{\omega}_k, \lambda_i, x)P(\tilde{\omega}_k|\lambda_i, x) \end{aligned} \quad (2)$$

where $P(\tilde{\omega}_k|\lambda_i, x)$ is the prediction of the individual classifier. Substituting eqn. 2 into 1 we arrive at the following:

$$\begin{aligned} P(\omega|x) &= \sum_{i=1}^{C}\sum_{k=1}^{K} P(\omega|\tilde{\omega}_k, x, \lambda_i)P(\tilde{\omega}_k|x, \lambda_i)P(\lambda_i|x) \\ &\approx \sum_{i=1}^{C}\left[\sum_{k=1}^{K} P(\omega|\tilde{\omega}_k, \lambda_i)P(\tilde{\omega}_k|x, \lambda_i)\right]P(\lambda_i|x) \end{aligned} \quad (3)$$

The last line of this equation the conditional error distribution, $P(\omega|\tilde{\omega}_k, x, \lambda_i)$, which is difficult to obtain, is approximated by its projection, $P(\omega|\tilde{\omega}_k, \lambda_i)$. The latter is simply an empirical distribution that can be obtained from the confusion matrix of the classifier on a validation subset of the training data.

The essence of equation 3 is that the prediction of each classifier is weighted in accordance to the error distribution over the classes.

Practical implications of this procedure involve multiplying the classifier scores with the empirical error distribution to obtain the corrected score that takes into account the certainty of the classifier about a particular class. Note that this combination framework should not significantly affect the output of a classifier which is in general accurate, since its confusion matrix will be close to identity. On the other hand, if a classifier systematically mistakes samples of, say, class 1 for samples of class 2, the prediction of the classifier about class 1 will be biased towards class 2 in proportion to the number of mistakes made on the validation set. While a good classifier should not be affected by such an operation, in combination with others more weight is given to class 2 and it is left to other classifiers to disambiguate this situation. It results in a "per-class" weighting scheme, rather than the traditional "per-classifier" paradigm.

Additionally, each classifier is weighted by the term $P(\lambda_i|x)$, which can express external knowledge about the instantaneous performance of each classifier. For instance, if some of the features are not present, the corresponding probabilities can be set to 0 and their outputs subsequently ignored in making the combined decision.

This model establishes a general framework for classifier combination, from which a variety of different combination strategies can be derived. In particular, Tax *et. al.*, [10], present a framework in which sum and product rules are formally justified. Our framework is fully compliant with their work in that we implicity allow for critic-based (induced by $P(\lambda_i|x)$) and error-corrected (induced by $P(\omega|\tilde{\omega}, x, \lambda_i)$) sum and product combination schemes.

3 Experiments

In section 3.1 we first explore the combination scheme on a synthetic data set, while in section 3.2 we apply the scheme

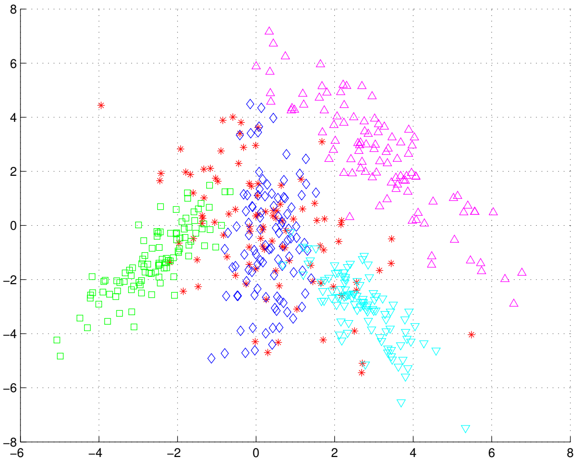

Figure 2: Example of the training data used for a single trial

to a real data set collected in an office environment by our automated surveillance system.

3.1 Evaluations on Synthetic Data

In the experiments on synthetic data we measure the performance of the combination scheme on a set of *trials*. A *trial* consists of the following:

1. Generate *training*, *validation* and *test* data.
   The data sets are drawn from five 2-dimensional Gaussian densities with random means and covariances. The means are drawn from normal distribution $\mathcal{N}([0,0]^T, 3*I)$, while covariances are samples drawn from the Wishart distribution $\mathcal{W}(I,3)$ (unit covariance, 3 degrees of freedom). One such data set is shown in fig. 2.

2. Train base classifiers
   We train 2 5-class base classifiers - one with equal weights, C_e, and the other with confidence weighting C_c. In the interest of fairness the equal weight classifier C_e is trained on both *training* and *validation* sets, while the classifier C_c is only trained on *training* data. For C_c the validation set is only used to calculate confusion matrices. In the limit of infinite data the performance of these classifiers should tend to optimal Bayes.

3. Generate classifier ensembles
   The two ensembles, E_e and E_c, are generated by random perturbation of the means of the base classifiers. We use 5 feature classifiers for each of the 2 base classifiers while not including C_e and C_c in the ensembles themselves.

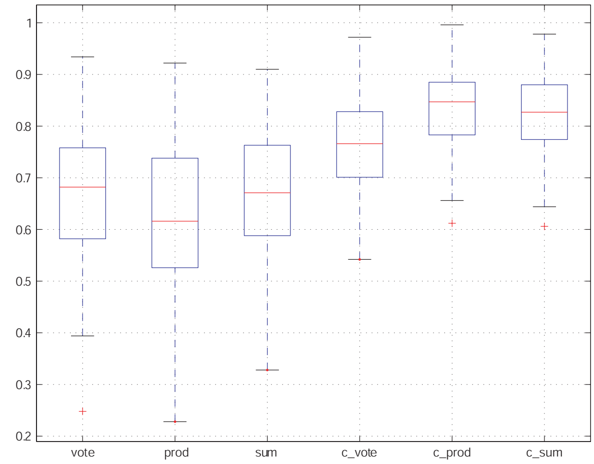

Figure 4: Performance of the Bayesian combination scheme as compared to the traditional combination techniques. The first three boxes correspond to voting, product and sum rules. The next group of three corresponds to the same rules but applied to the weighted classifier scores. Each box represents the data within 25th and 75th percentile, with the mean marked as a horizontal line. The whiskers of the plot relate to the variance of the accuracy of the classifiers over the set of trials

Table 1: Summary of the results presented in figure 4. The table shows that the weighted combination rules consistently outperform the vanilla Vote, Product and Sum rules

	Equal		**Weighted**	
	Mean	**Variance**	**Mean**	**Variance**
Vote	0.6696	0.0171	0.7621	0.0088
Product	0.6257	0.0223	0.8373	0.0065
Sum	0.6749	0.0161	0.8272	0.0069

4. Calculate error rates
   Evaluation of the performance of both ensembles is done on the *testing* data.

Figure 3 shows ROC curves for the three combination rules - voting (a), product (b) and sum (c). In all cases the classifier ensemble using the weighting scheme derived from the approximation to the error function shows better performance than the ensemble based on the classifier using both training and validation data sets.

In figure 4 and table 1 we report average error rates over 100 trials along with their variances. The error rates are effectively averaged over possible choices of classifiers.

3.2 Experiments with HID data

In the second set of experiments we apply the combination technique to surveillance data recorded automatically

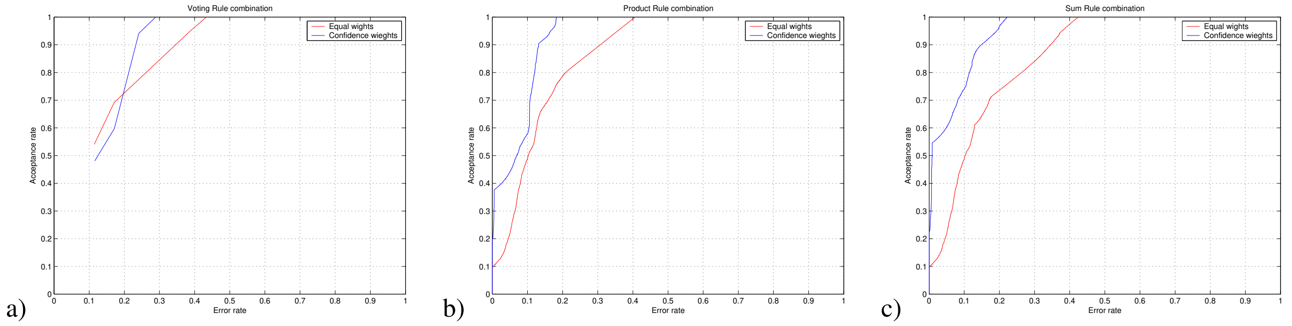

Figure 3: ROC curves for a single trial showing three different classifier combination strategies with and without confidence weighting. a) voting combination rule; b) product rule; c) sum rule. In all cases the top curve is generated by a weighted rule

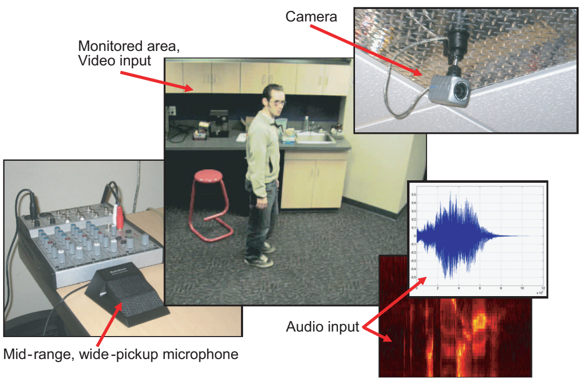

Figure 5: MMHID system for on-line multi-modal human identification.

from our on-line multi-modal human identification system (MMHID). The data is collected from 6 people in the office environment over the course of about 3 weeks. The surveillance setting implies that no manual intervention is employed to clean up the data, that is, if some features are incorrectly detected, they are not removed from the data set. That reflects the true circumstances under which the system should be deployed.

The system, shown in figure 5, receives input from a statically mounted video camera and microphone. The classification proceeds on the features extracted from the camera and microphone in real time. We stream both audio and video channels to disk in threads with the highest priority, while tapping into the streams to extract data for each individual classifier. We process audio and video channels independently.

3.2.1 Audio Features

For voice classification we use the well known MFCC representation. First, 40 MEL Frequency Coefficients (MFCs) captured at 100Hz from overlapping sliding windows. Then they are converted to Cepstral Coefficients (MFCCs). Only the frames with majority of power focused into the lower half of the Mel-scale frequency spectrum (lower 20 MFCs) and having significant energy are tagged as voice features to be included in later classification steps.

After obtaining sufficient amount of the audio samples we train a Gaussian Mixture-based classifier on the collected set of Mel-Scale Cepstral Coefficients. In our experiments we use an 8-component mixture to model a person.

3.2.2 Video Features

The extraction of video features begins with detecting a person in the view of the camera. To detect a person we apply a set of simple rules, such as: presence of significant motion over at least 3 frames; sufficient and stable illumination conditions; and appropriate aspect ratio of the enclosing bounding box. If such conditions are met, we trigger recording.

To extract the video features we first perform background subtraction and sum the resulting mask vertically. In the resulting histogram of the non-zero values we find the peak and expand it outwards until a sum below a chosen threshold T is encountered. We use this range as an estimate of the horizontal position and extent of the bounding box. We found this estimate to be more robust, as it is less likely to include shadows that a person might drop on furniture and walls.

When the subject has later left the camera's field of view, recorded signals are automatically timestamped and entered into an SQL database for later processing.

3.2.3 Face

For each frame in the video stream, we perform face detection [3] over a bounding box region. If a person is present, we scan the box containing the object of interest for a face and, if one is found, a smaller patch enclosing only the face is extracted from the image for classification.

With a labeled dataset of faces of K people we train K one-vs-all second order polynomial SVM classifiers using

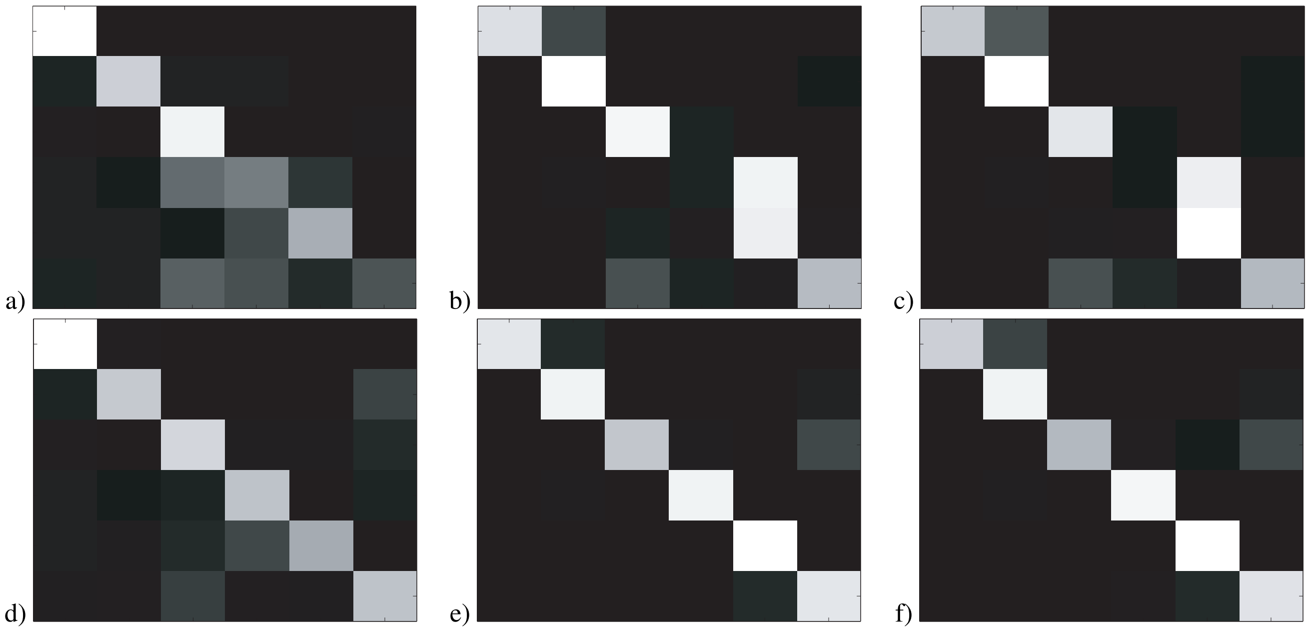

Figure 6: Confusion matrices calculated on the test set. a)-c) vote, product and sum rules applied to the outputs of the classifiers directly. d)-f) same with error-weighted outputs. Element of the matrix $C_{i,j}$ contains a number of times a person i was classified as person j in the data set.

the SVMFu package [8].

3.2.4 Height

Having a calibrated camera system we determine the person's height from the bounding box in the image. Calibration gives us access to intrinsic parameters, A (focal length and principal point), as well as extrinsic parameters R and t, which define rotation and translation of the camera coordinate system with respect to some known position in the scene. With these, the imaging relation for the camera system is given by the following:

$$\tilde{m} = A\,[R|t]\,M \tag{4}$$

where M is the 3D coordinate of the imaged point, and $\tilde{m}$ is the vector of homogeneous coordinates of the point in the image plane.

To invert this projective relation we use additional constraints which assume that a) the person is standing on the ground plane co-located with the (X, Y) plane of the global coordinate system; and b) the person's body is vertical.

We compute the height of the person by solving a system of linear equations for two points - at the top and at the bottom of the bounding rectangle:

$$M = R^{-1}\left(A^{-1}\tilde{m} - t\right) \tag{5}$$

subject to the ground plane constraints[1].

From these estimates we estimate a single gaussian density per person, which is subsequently used for classification.

3.2.5 Clothing

We capture individual clothing preferences by way of separate upper and lower body histograms, under the assumption that, within a given day, individuals do not change their clothing. The histogram is computed from Hue and Saturation components of the image in the HSV color space. We allocate 32 and 24 bins to the H and S components, respectively. Histograms are then labeled according to subject, and averaged by day so that for each day there are single upper and lower average histograms for each user. Collections of histograms for a given user are finally clustered using K-means ($K = 5$), to give a clothing model consisting of K prototypes.

These prototypes are later compared to test histograms using normalized cross-correlation during classification.

3.2.6 Results

In the combination of the classifier outputs we use a simple binary critic, $P(\lambda|x)$ (see eqn. 1. For each frame of the

[1] We also apply the bounding box correction due to camera roll, as well as ground plane correction due to estimation errors in the image-ground homography parameters.

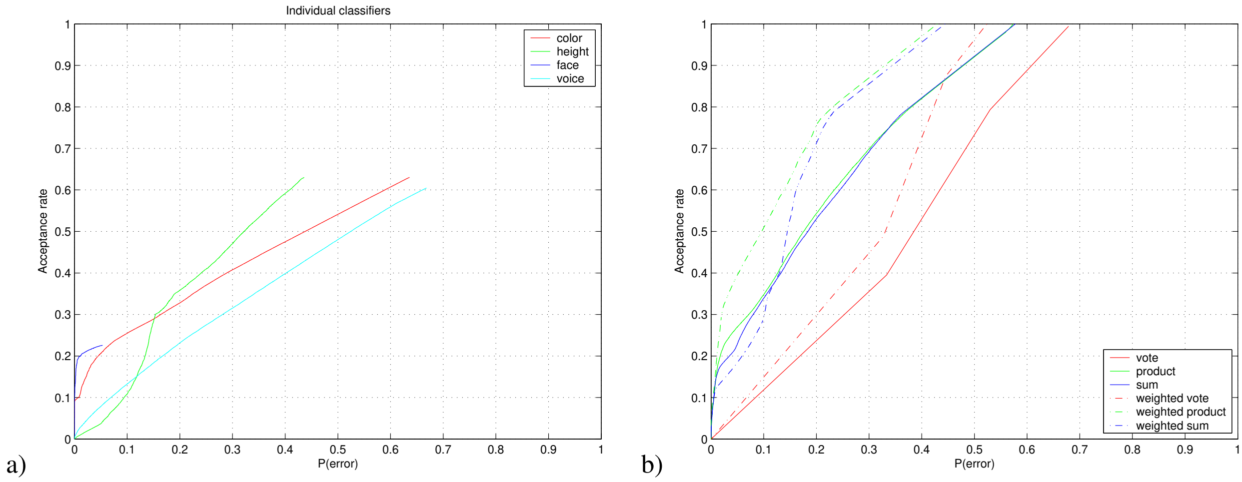

Figure 7: Multi-class ROC curves for individual and combined classifiers. The acceptance rate is computed as a proportion of accepted in the complete set of frames. a) Individual classifiers; b) combined classifiers. Combination rules include both error-weighted and not weighted ones.

video sequence we assign probability 0 to classifiers which do not observe their feature in the current frame. The rest of the classifiers are weighted equally.

The results of running the vote, sum and product combination rules for direct and error-weighted classifier combination are shown in figures 6 and 7.

It can be seen from the figure 6 the weighted combination schemes achieved superior classification performance as compared to direct combination rules. The figure shows confusion matrices of the resulting classifiers. True person's identity is arranged along the vertical columns. The count of classifier decisions is shown in rows. In each figure a brighter color corresponds to a higher count. It is clearly seen that the bottom row of matrices represents less confusions about people identities in the classifiers derived from the error-weighted scheme.

More precisely, the performance of the combination schemes is illustrated by their ROC curves. The multi-class version of these curves shows acceptance rate of the classifier as a function of the error rate. Thus, the value of the curve at 100% acceptance shows the error rate of the classifier on the full data set (classification regime). A verification regime of the classifiers can be evaluated from this curve by fixing the error rate and comparing the corresponding acceptance rates.

It is illustrated in the figure 7a) that even though the face classifier in its full acceptance mode can achieve a very high accuracy, it can only deliver this performance on a relatively small fraction of the full video sequence. In contrast, height, color and voice classifiers are relatively poor, but have observations in about 60% of the whole sequence. Combining these classifiers allows us achieve performances shown in figure 7b). It needs to be emphasized that the resulting classifier produces the classification decision for almost all frames of the video sequence using whatever information is available at the moment.

Figure 7b) shows relative performances of different combination schemes. On this data set, as well on the synthetic set from the previous section error-weighted combination schemes consistently outperform the direct combination approaches.

4 Discussion and future work

In this paper we introduced an error correcting classifier combination technique and applied it to two data sets - a synthetic set and a set of multi-modal data collected from a automated surveillance system. Experiments on synthetic data illustrated that application of the technique resulted in higher accuracy classifier system with smaller variance than the well known direct techniques.

Application of this technique to the surveillance data set allows us to improve the recognition accuracy on a sample basis and deploy the combination in an on-line setting. In this setting we are able to give a prediction of the person's identity in almost every frame of the input video sequence, while each individual classifier can only do it for a fraction of the data set. It is clear that the performance can be much improved with temporal integration, but this is outside the scope of the paper. Our future work will include using the technique presented in this paper in combination with temporal integration as well as direct estimation of the conditional error of eqn. 3.

References

[1] Chiraz BenAbdelkader, Ross Cutler, and Larry Davis, *Stride and cadence as a biometric in automatic person identification and verification*, FG, 2002.

[2] J. Bilmes and K. Kirchhoff, *Directed graphical models of classifier combination: Application to phone recognition*, Intl. Conference on Spoken Language Processing (Beijing, China), 2000.

[3] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio, *Categorization by learning and combining object parts*, Advances in Neural Information Processing Systems 14 (Vancouver, Canada), 2002, pp. 1239–1245.

[4] J. Kittler, Y. P. Li, J. Matas, and M. U. Ramos Sánchez, *Combining evidence in multimodal personal identity recognition systems*, Intl . Conference on Audio- and Video-Based Biometric Authentication (Crans Montana, Switzerland), 1997.

[5] J. Kittler, G. Matas, K. Jonsson, and M. Sánchez, *Combining evidence in personal identity verification systems*, Pattern Recognition Letters **18** (1997), no. 9, 845–852.

[6] J. Nam, E. Cetin, and Tewfik, *Speaker identification and video analysis for hierarchical video shot classification*, Int. Conf. Image Processing (Santa Barbara, CA), 1997.

[7] E. Pekalska, R. Duin, and M. Skurichina, *A discussion on the classifier projection space for classifier combining*, 3rd International Workshop on Multiple Classifier Systems (Cagliari, Italy), Springer Verlag, 2002, pp. 137–148.

[8] Ryan Rifkin, *SVMFu package*, http://five-percent-nation.mit.edu/SvmFu/index.html, 2000.

[9] Arun Ross and Anil K. Jain, *Information fusion in biometrics*, Pattern Recognition Letters, Vol. 24, Issue 13, pp. 2115-2125 **24** (2003), 2115–2125.

[10] David M. J. Tax, Martijn Van Breukelen, Robert P. W. Duin, and Josef Kittler, *Combining multiple classifiers by averaging or by multiplying?*, Pattern Recognition **33** (2000), 1475 – 1485.

[11] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, *Face recognition: A literature survey*, Tech. Report CAR-TR-948, University of Maryland, 2000.