# On Probabilistic Combination of Face and Gait Cues for Identification

Gregory Shakhnarovich Trevor Darrell
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
200 Technology Square, Cambridge MA 02139
{gregory,trevor}@ai.mit.edu

## Abstract

*We approach the task of person identification based on face and gait cues. The cues are derived from multiple simultaneous camera views, combined through the visual hull algorithm to create imagery in canonical pose prior to recognition. These view-normalized sequences, containing frontal images of face and profile silhouettes, are separately used for face and gait recognition, and the results may be combined using a range of strategies. We discuss the issues of cross-modal correlation and score transformations for different modalities, present the probabilistic settings for the cross-modal fusion. and explore several common fusion approaches. The effectiveness of various strategies is evaluated on a data set with 26 subjects. We hope that the discussion presented in this paper may be useful in developing further statistical framework for multi-modal recognition.*

## 1. Introduction

Visual recogintion of individuals from multiple arbitrary views is an important task for many applications. Perceptual interfaces for intelligent environments, visual surveillance and activity monitoring, and covert security and access control can all benefit from recognition at a distance. These applications usually can not presume that users will present themselves in a canonical pose or be close to the camera. For optimal performance, a system should incorporate as many observations as are available, and extract as many informative cues as is possible.

We have developed a system for recognition from multiple video streams, based on a canonical view rendering technique applied to face and gait recognition algorithms. Given a set of images from multiple cameras, we reconstruct virtual views in canonical pose: frontal for face, and profile for gait. In [10] it was demonstrated that synthetic views rendered with a visual hull [8] improved recognition significantly over results with unnormalized imagery, from 52% to 90%. It was also shown that the combination of face and gait cues provided slightly better recognition results than either modality alone, using a simple average of classifier outputs across modality and over time.

In this paper we investigate different approaches to classifier combination for face and gait recognition, and demonstrate both improved performance and better statistical justification for the integration step. First, we compare the performance of several common data fusion strategies on our task, and develop statistical interpretations of each. Following the theoretical framework presented in [6], we compare MAX, MIN, MEAN, and PRODUCT rules for combining classifier outputs. We assess the underlying assumptions for each model, and empirically evaluate which ones are appropriate for the task of face and gait integration.

Second, we explore the effect of early versus late temporal integration for instantaneous features. Since gait recognition is performed over an entire sequence of data, no temporal integration occurs. But face cues (images) are generated per set of input frames, and the order of recognition and temporal integration is arbitrary. With late integration, classification is performed on each frame and the resulting scores or probabilities are combined. With early integration, the features themselves are combined over time before being passed to the classifier. In both cases, one can consider the range of combination schemes mentioned above.

We use existing algorithms for face and gait recognition, based on a subspace-per-user eigenfaces technique [11], and a technique for gait recognition based on spatio-temporal motion sequence matching [7]. Face and gait are appropriate features to use on multi-view sequence data, since they capture apparently independent characterisitcs of users. Face cues are derived from the relatively detailed instanteous appearance of the face surface, while gait cues are obtained from coarse body shape as it moves over time.

In the remainder of this paper, we will review relevant prior work, our scheme for recogntion from synthetic canonical views, and the face and gait recognition algorithms used in our system. We will then discuss different

integration strategies used for combining face and gait data, and for combining face data over time. Finally, we will present the results of experiments involving 26 subjects, and conclude with a discussion of the implications of these results and avenues for future work.

## 2. Fusion for Recognition

The general topic of sensor fusion for pattern recognition has a substantial literature. For the specific task of biometric recognition, a variety of approaches are possible, a few of which we mention here. Voting schemes, such as those used in [3] for integrating face, lip, motion,and voice, ignore all non-winning individuals in each modality. Rank ordering approaches have been used by several authors, including [2] for face/voice integration and [13] for recognizing commands using speech and pen gesture. In [5] a framework for integrating multiple biometric cues in a large database application was provided, where a portion of the cues were used for retrieval and the remainder for verification, and also explored late vs. early fusion in the context of fingerprint recognition. In [6] a theoretical framework was developed for combining independent classifiers, and different sets of simple assumptions were shown to lead to a range of commonly used combination heuristics. We followed in our experiments.

## 3. Integrated recognition by face and gait

We shall briefly review the view-normalized recognition system, with which our face and gait data are obtained, and the face and gait classifiers involved.A more detailed description can be found in [10].

### 3.1. General overview

The system is based on $K$ monocular cameras $c_1, \ldots, c_K$, synchronized by hardware and calibrated in a coordinate system in which the $XZ$ plane coincides with the ground plane. In our immplementation $K = 4$. At each time $t$, $c_i$ provides data consisting of a "raw" image $I_i^t$ and the silhouette $S_i^t$ computed by means of foreground/background detection (Figure 1, (a)). Let $F^t$ be the collection of all $K$ inputs at time $t$. From $F^t$ we construct the visual hull [8] – a geometric model constructed from the intersection of the 3D cones defined by the $K$ silhouettes and centers of projection of the cameras. The visual hull is approximated by a triangular mesh $H^t$ as shown in Figure 1, (b). The projection of $H^t$ to an arbitrary image plane defines a *synthetic silhouette* of the object. Texture can then be mapped onto the silhouette, based on $I_i^t$ and on the desired vantage point, thus producing a *synthetic view* (bottom row of Figure 1(b)). It remains to choose the viewpoint optimal for each of the recognition modalities. Such viewpoint is defined in terms of the motion trajectory of the person.

The trajectory of the moving person can be estimated by fitting a curve to the observed locations of the centroid of $H^t$ for each $t$. Assuming linear motion of the subject, one can find a least-squares fit. In a more general case, a Kalman filter can be applied to the measured locations [10]. Estimation of the tangent to the trajectory, together with the ground plane to which the motion is assumed to be roughly parallel, establishes canonical axes and allows us to produce synthetic views of the person from a desired vantage point relative to his/her body orientation.

### 3.2. View-normalized face recognition

We use eigenfaces approach [11] for recognition of view-normalized faces. For each subject $w_k$ in the database, it computes the distance from the presented image $\mathbf{f}$ to the subspace spanned by $M$ principal components $\mathbf{B}_k$ of the training set of faces for that class. The input is assumed to be a frontal facial image, which makes view-normalization crucial. The score of $w_k$ given the input $\mathbf{f}$ is computed as

$$D_f(w_k|\mathbf{f}) = \|\mathbf{f} - \mathbf{f}^T\mathbf{B}_k\|,$$

and the classes are ranked by increasing score.

Under the mild assumptions of upright body posture and fronto-parallel motion[1], the expected position of the face is in the top portion of $H^t$, facing the direction of motion within a small angle. Note that the scale is known from the distance of the virtual image plane from $H^t$. The face is then sought in all relevant subimages using a fast face detector [12]. The detected "box" is resized to a base size, which in our experiments is $22\times22$.

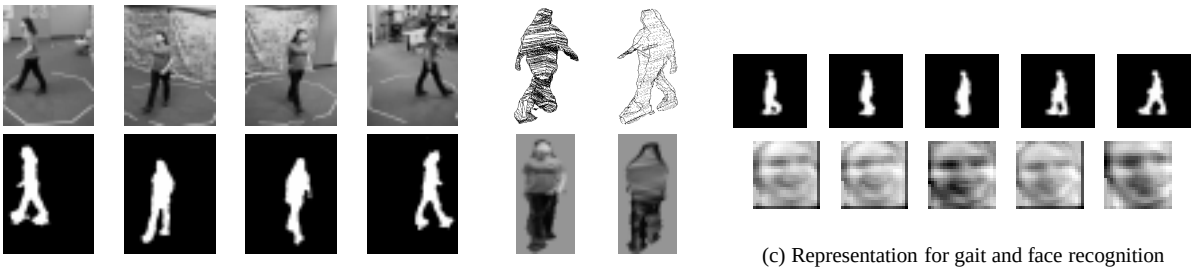
At a given time $t$ we can have more than one face detections, corresponding to different spatial angles with respect to the estimated motion direction. For simplicity, let us assume that we pick one of the faces, say, that with the highest score with respect to a certain criterion (such as highest estimated degree of frontality)[2]. Sometimes, however, no faces are detected due to visual hull inaccuracies, face detector failure or momentary occlusion of the face. In the worst case, in which not a single face is detected for the whole sequence, the face modality can not be used for recognition.

### 3.3. View-normalized gait recognition

We use a simple and robust algorithm for matching spatiotemporal sequences recently proposed in [7]. In this al-

[1] These assumptions may be relieved by performing a more exhaustive search at $t = 1$ and tracking the face in the subsequent frames.

[2] Alternatively, we may use all of the available face detections for classification.



(a) "Raw" input (b) Model (c) Representation for gait and face recognition

**Figure 1. Examples of input(a) and output(b) of visual hull and its use for constructing synthetic input for face and gait classifiers(c)**

gorithm, each frame containing the silhouette of a walking subject is divided into a small number of fixed regions, and moments of the foreground pixels are computed in each region. The means and standard deviations of these moments, along with the centroid of the entire silhouette, comprise the feature vector $\gamma$ for a given sequence of silhouettes $\langle S^1, \ldots, S^t \rangle$. This representation relies on the geometry of the projected silhouettes and is therefore view-dependent. However, the visual hull representation allows us to create a sequence of synthetic silhouettes[3]from any desired view, making view-independent gait recognition possible [10].

The training data for the algorithm consists of a collection of labeled profile sequences $\{\langle \mathbf{S}_j, l_j \rangle\}$. When a test sequence $\mathbf{S}$ is presented to the algorithm, the score for each class is computed as

$$\Gamma(w_k|\mathbf{S}) = \min_{l_j = w_k} \|\gamma(\mathbf{S}) - \gamma(S_j)\|$$

### 3.4. Score transformation

Having obtained the score for each model given the observations in each modality, one generally can not directly combine these scores in a statistically meaningful way. The gait classifier produces scores which are not direct estimates of the posterior, but rather measures of the distance between the test example the best matching reference feature vector of the $k$th person. The face classifier, in addition to the distance between a face and the $k$th eigenspace provides an estimate of the likelihood of that face under $k$th model. These scores, with quite different ranges and distributions, must therefore be transformed. In order to justify the scored-based decision statistically, such a transformation must assume a monotonic growth in score of a model given the data as a function of the posterior probability of that model.

[3]In fact, we can produce two synthetic silhouettes for each frame, as viewed from both sides of the body, and combine the resulting feature vectors or the recognition results

A number of heuristics for score transformation have been proposed in the literature [1, 4]. Below we attempt to establish a generalized approach to score transformation for classifiers that output distances in some metric space, with no analytic form of the class-conditional or the posterior available. For the sake of clarity, we shall refer to the gait classifier; however, it should be stressed that it applies to a more general case. In particular, the score of the face classifier can be transformed in a similar way.

We assume a probability distribution over the scores assigned to the **correct** labels – essentially, the distribution of the distances between two representations of a person[4], and try to model this distribution $\hat{P}(\cdot)$. Examination of the empirical distribution over the observed scores for both correct and incorrect labels (Figure 2) suggests that this approach is valid. The proposed estimate of the posterior can then simply be

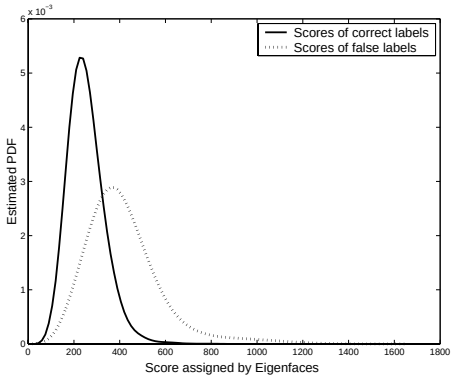
$$P_{\text{gait}}(w_k|\mathbf{S}) = P_{\text{gait}}(w_k|\Gamma(\mathbf{S})) = \hat{P}(\Gamma(\mathbf{S})$$

(normalized to sum to unity over all $w_k$). In practice this means fitting a function $T$ to the empirical distribution, and treating $T(\Gamma(\mathbf{S}))$ as the estimate of the posterior.

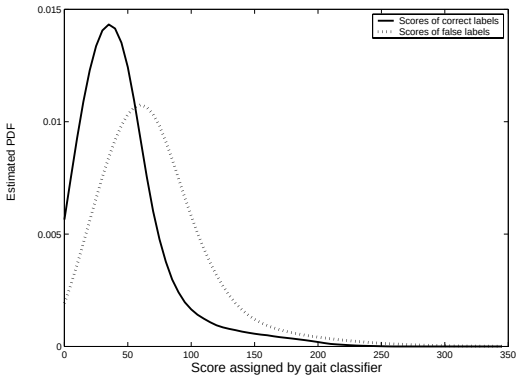
The solid line in Figure 3 shows the empirical probability density computed over the training data, and the logistic function $1/\left(1 + e^{ax/C}\right)$, which we used as the transformation function. The parameter $a$ controls the slope, $C$ is a normalization constant that ensures that the argument of the logistic function is contained in $[0, 1]$. This constant may be found analytically, through parametric regression, or empirically (as it was in our case).

Clearly, when a specific probabilistic model of the classification process is available, including some estimate for the posterior, this model can replace the proposed estimation scheme.

[4]The implicit assumption that this model is the same for each class is somewhat simplistic, and may be lifted in the future work




(a) Face




(b) Gait

**Figure 2. Distribution of the scores assigned to correct (solid lines) and false (dotted lines) labels by face and gait classifiers**

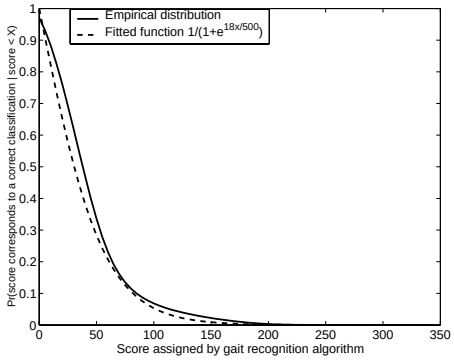



**Figure 3. Modeling the distribution of the scores for gait recognition algorithm. The empirical distribution is well approximated by a logistic function**

## 4. Temporal fusion

One of the fusion problems in the presented framework is related to multiple available images of a person's face. Ii is an instance of a general problem, where a set or a sequence of observations, all belonging to the same domain, are known to belong to the same class. We have a choice between two options: "early" and "late" fusion. In both cases, we assume statistical independence between images in the set, an assumption clearly incorrect in the case of temporally adjacent face images with dynamic expression etc.

Early fusion, i.e. fusion on the *sensor level*, consists of combining the observations (in our case separate face images ) and mapping them into a single data point to be classified. This can be achieved, for example, by computing the mean face in the input set and presenting it to the recognition algorithm. A more sophisticated approach is to estimate a probabilistic model of the input set and compare it directly to the learned models of the database subjects [9].

Alternatively, one can treat all the images as separate independent data, and perform late fusion, at the *decision level*. This can be done, for example, using one of the relevant combination rules discussed in Section 5. In our experiments, we found that when working with the scores as opposed to posteriors, the distance of the mean face to the subspace provides a good way of classifying the whole set of face images. When posteriors or likelihoods are given, their product acts as the optimal combination rule.

## 5. Cross-modal integration

Early fusion of data in domains so different as silhouettes and face images is an important subject of future research. However, at this time it is not clear how to combine the information at the sensor level. Thus we resort to decision-level fusion. Previously, an *ad-hoc* combination rule was used, namely, averaging the normalized scores of the two classifiers and choosing the first ranking label. However, this rule is not optimal for the case of highly uncorrelated (or independent) classifiers. In [6] a number of common combination schemes were given a theoretical justification, which we shall discuss below with regard to our recognition domain.

### 5.1. Combination rules

Let the feature input to the $j$-th classifier, $j = 1, \ldots, R$ be $\mathbf{x}_j$, and the winning label be $h$. We assume a uniform prior across all classes (identities) and shall omit it from the

formulae, but the rules allow for the incorporation of any knowledge about priors.

**PRODUCT rule**:$h = \text{argmax}_k \prod_{j=1}^{R} P(w_k|\mathbf{x}_j)$. This rule is derived under the assumption of conditional statistical independence of the representations:

$$P(\mathbf{x}_1, \mathbf{x}_2|[F^1, \ldots, F^N]) = P(\mathbf{x}_1|[F^1, \ldots, F^N])P(\mathbf{x}_2|[F^1, \ldots, F^N]).$$

It is important to evaluate the validity of this assumption. In the absence of a probabilistic model for gait features or the gait classifier, all we can examine is the amount of correlation in the data, and treat the lack of correlation as an evidence of independence. Figure 4 demonstrates that the features (on the left - from left to right and from top to bottom, gait features and the pixels of the face images) are correlated much stronger within each modality than across modalities. The correlation that does exist can be in pat explained by the parameters such as gender, which is obviously correlated with face appearance and, in light of the success of the same gait classification approach when applied to gender classification, with our gait features.

**MEAN (sum) rule**: $h = \text{argmax}_k \sum_{j=1}^{R} P(w_k|\mathbf{x}_j)$. This rule, derived from PRODUCT, is reported to be the winner in [6]. It is most applicable when a high level of noise and/or high ambiguity in the classification problem cause the posterior estimated by a classifier not to deviate much from the prior.

**MAX rule**: $h = \text{argmax}_k \max_j P(w_k|\mathbf{x}_j)$. This rule approximates the mean by the maximum of the posteriors.

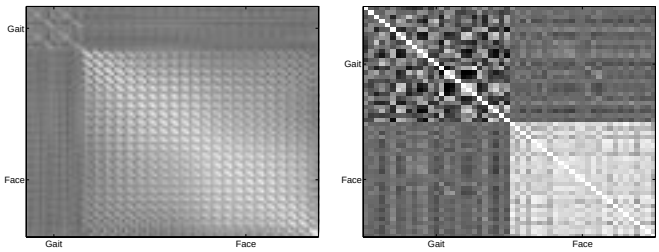
**MIN rule**: $h = \text{argmax}_k \min_j P(w_k|\mathbf{x}_j)$. An approach is to bound the product from above by the minimum.

**MAJORITY rule**: this rule chooses the class with the highest number of hard classifications. It is irrelevant here since we have only two classifiers. The weighted majority rule is not appealing either, since the individual classifiers exhibit very similar error rates, and it is not clear why one would prefer one of them over the other.

## 6. Experiments

We collected 206 data sequences of 26 people walking; the number of sequences per person varies from 2 to 14. For 11 of the subjects the data was collected on two separate days about 3 months apart. Lengths of the sequences range from 9 to 23, with an average of 15 frames, at 13 frames per second. Main results are shown in Figure 5. All reported results were computed by leave-one-out cross validation: each sequence was classified based on the rest.

The baseline single-modality classifiers have an accuracy of **68%** (gait) and **57-73%** (face, depending on the scheme for integration of the faces). Fusion on the sensor level (working with the face averaged over a sequence) produces a baseline face recognition rate of **69%**. The best combined accuracy using mean faces was **85%**. Our feeling is that due to the noise and misalignment, we lose information by performing early integration for face recognition.




(a) Correlation in the input (b) Correlation in the scores

**Figure 4. Correlation of the features (a) and the scores (b) for gait recognition. The cross-modal correlation is clearly much weaker than the correlation between features in the same modality.**

There are two main degrees of freedom in choosing the combination strategy. First, one has to choose the rule for combining multiple faces. Second, one has to choose the rule for combining the modalities. In our experience, the better performing combination rules – PRODUCT and MEAN – were robust to the changes in temporal furion of faces (compare the

Before starting the experiments with the rules discussed above, we tried a simpler rule which requires both classifiers to agree on a label for a test example; otherwise, the example is rejected. Since the classifier decisions appear to be uncorrelated (Figure 4), we expect the accuracy to be close to the the product of the individual accuracies, which is 45%. The observed accuracy was indeed **49%**.

The best performance was achieved by the PRODUCT rule:**89%** accuracy. When the quantities combined are estimates of the posterior distribution, and under the independence assumption, this rule can be shown to be equivalent to the likelihood ratio hypothesis test, when the joint likelihood of the combined data is considered under different models, and equal priors are assumed.

We believe that the main reason for the poor performance of the MIN and MAX rules is the high degree of overlap of the distributions of correct and incorrect scores for the classifiers (see Figure 2). Both rules rely on order statistics and are likely to suffer from the noise in score assignment more than the more robust MEAN and PRODUCT.

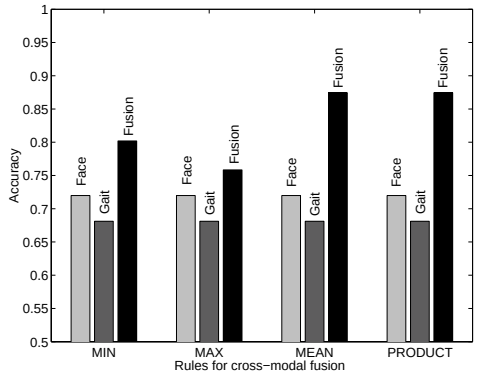



**Figure 5. Result of various cross-modal fusion rules, with faces integrated with PRODUCT rule**

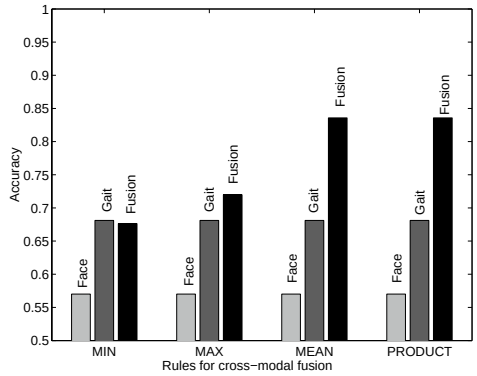



**Figure 6. Cross-modal fusion, with faces integrated by MIN rule**

## 7. Conclusions

We have developed a probabilistic approach to combining visual cues for human recognition, as well as for using multiple instances of face classifications, and demonstrated its performance on the example of integrated face and gait recognition. A number of previously proposed combination rules have been empirically compared. While the combination improved the classification accuracy of the system in almost all cases, the best performance was obtained by a classifier that uses product of the posterior probabilities estimated from different modalities. This classifier achieved an improvement of 15% in the leave-one-out test performance.

Our study highlights the importance of a careful choice of the whole combination strategy. Some of the straightforward rules, such as MIN, performed poorly in our experiments and at times proved harmful (Figure 6). Score transformation appeared to be another important issue.

Interesting future work includes extension of this bimodal recognition scheme to additional modalities (color distributions, voice, activity patterns), thus making more complex rules relevant. Finally, the combination strategy remains largely an *ad-hoc* endeavor. Exploring the decision-level fusion in a Bayesian context, with regard to the estimated distributions of the correct and incorrect labels, may lead to a more theoretically sound understanding and design of this process.

## References

[1] B. Achermann and H. Bunke. Combination of classifiers on the decision level for face recognition. Technical Report IAM-96-002, Insitut fur Informatik und angewandte Mathematik, Universitat Bern, Bern, 1996.

[2] B. Brunelli and D. Flavigna. Personal identificaion using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[3] U. Dieckmann, P. Plankensteiner, R. Schamburger, and B. Froeba. SESAM: A biometric person identification system using sensor fusion. *Lecture Notes in Computer Science*, 1206:301–??, 1997.

[4] R. Duda, P. Hart, and D. Stork. *Pattern classification*. John Wiley & sons, New York, second edition, 2001.

[5] L. Hong and A. Jain. Multimodal biometrics. In A. Jain, R. Bolle, and S. Pankanti, editors, *Biometrics: Personal Identification in Networked Society*. Kluwer, 1999.

[6] J. Kittler, M. Hatef, R. Duin, and J. Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar. 1998.

[7] L. Lee. Gait dynamics for recognition and classification. Technical Report AIM-2001-019, MIT AI Lab Memo, Sept. 2001.

[8] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. In K. Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, Annual Conference Series, pages 369–374. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

[9] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proceedings of European Conference on Computer Vision*, 2002. to appear.

[10] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated Face and Gait Recognition From Multiple Views. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, Lihue, HI, Dec. 2001.

[11] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuro Science*, 3(1):71–86, 1991.

[12] P. A. Viola and M. J. Jones. Robust real-time object detection. Technical report, COMPAQ Cambridge Research Laboratory, Cambridge, MA, Feb. 2001.

[13] L. Wu, S. L. Oviatt, and P. R. Cohen. Multimodal integration - a statistical view. *IEEE Transactions on Multimedia*, 1(4):334–341, 1999.