FUSION OF GAIT AND FACE FOR HUMAN IDENTIFICATION

Amit Kale, Amit K. RoyChowdhury and Rama Chellappa[*]

Center for Automation Research
University of Maryland at College Park
College Park MD 20742 USA

ABSTRACT

Identification of humans from arbitrary view points is an important requirement for different tasks including perceptual interfaces for intelligent environments, covert security and access control etc. For optimal performance, the system must use as many cues as possible and combine them in meaningful ways. In this paper we present fusion of face and gait cues for the single camera case. We employ a view invariant gait recognition algorithm for gait recognition. A sequential importance sampling based algorithm is used for probabilistic face recognition from video. We employ decision fusion to combine the results of our gait recognition algorithm and the face recognition algorithm. We consider two fusion scenarios: hierarchical and holistic. The first involves using the gait recognition algorithm as a filter to pass on a smaller set of candidates to the face recognition algorithm. The second involves combining the similarity scores obtained individually from the face and gait recognition algorithms Simple rules like the SUM, MIN and PRODUCT are used for combinining the scores. The results of the fusion are demonstrated on the NIST database which has outdoor gait and face data of 30 subjects.

1. INTRODUCTION

Identification of humans from arbitrary view points is an important requirement for different tasks including perceptual interfaces for intelligent environments, covert security and access control etc. Different modalities can be used for identification based on the distance of the individual from the camera. If the person is far away from the camera, it is hard to get face information at a high enough resolution for recognition tasks. However when available, it yields a very powerful cue for recognition. A modality which can be detected and measured when the subject is far away from the camera is human gait or the style of walking. For optimal performance, the system must use as many cues as possible and combine them in meaningful ways. Information may be fused in two ways. The data available may be fused and a decision can be made based on the fused data (data fusion) or each signal/feature can be matched separately, using possibly different techniques and the decisions made may be fused (decision fusion).

The gait of a person is best reflected when he/she presents a side view (referred to in this paper as a canonical view) to the camera. Hence, most gait recognition algorithms rely on the availability of the side view of the subject. For doing face recognition, on the other hand it is desirable to have frontal views of the person's face. The most general solution to perform integrated face and gait recognition from arbitrary views would be to estimate 3-D models for face and gait. While there has been some progress in building 3-D models for faces [1], the problem of building reliable 3-D models for articulating objects like the human body still remains a hard problem. One way to exploit current recognition algorithms for frontal face and side gait without resorting to 3-D models is to synthesize canonical views, given arbitrary views of the person. In [2], Shakhnarovich et al. compute an image based visual hull from a set of monocular views which is then used to render virtual canonical views for tracking and recognition. Gait recognition is achieved by matching a set of image features based on moments extracted from the silhouettes of the synthesized probe video to the gallery. The visual hull is also used to render frontal face images. Eigen faces [3] are used for face recognition. In a later work, Shakhnarovich and Darrell [4] studied the fusion of face and gait cues for this multi-camera indoor environment.

In general the visual-hull approach for performing integrated face and gait recognition requires at least two cameras. In this paper we present experimental results for fusion of face and gait for the single camera case. We considered the NIST database which contains outdoor face and gait data for 30 subjects. In the NIST database, subjects walk along an inverted Σ pattern(see Figure 2. In one segment of the NIST database the subjects walk at an angle to the exact side-view. At the end of the walking path the person walks providing a nearly frontal view of his face to the camera. This final segment can be used for face-recognition. In [5], we presented the results of our view-invariant gait recognition algorithm for the single camera case on a database of thirteen people. In this paper we present the results of our view-invariant gait recognition algorithm in [5] on the NIST database. The algorithm is based on the planar approximation of the person which is valid when the person walks far away from the camera. In [6], results of an algorithm for probabilistic recognition of human faces from video was proposed and the results were demonstrated on the NIST database. We employ decision fusion which is a special case of data fusion (see Kokar et al. [7]) to combine the results of our gait recognition algorithm and the face recognition algorithm in [6]. We consider two fusion scenarios: hierarchical and holistic. The first involves using the gait recognition algorithm as a filter to pass on a smaller set of candidates to the face recognition algorithm. The second involves combining the similarity scores obtained individually from the face and gait recognition algorithms Simple rules like the SUM, MIN and PRODUCT are used for combinining the scores.

[*] Supported by the DARPA/ONR grant N00014-00-1-0908.

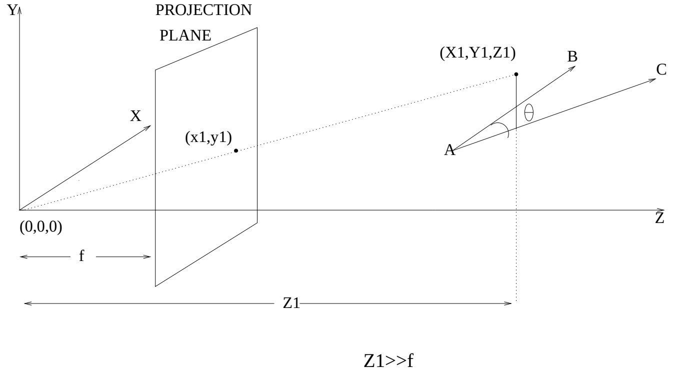


Fig. 1. Imaging Geometry

2. FRAMEWORK FOR VIEW-INVARIANT GAIT RECOGNITION

The imaging setup is shown in Figure 1. We assume that the person walks with a translational velocity $\mathbf{V} = [v_X, 0, v_Z]^T$ along the line AC at an angle θ to the canonical direction AB. Assuming that a high level motion detection module identifies segments of *approximately* constant heading directions and that we can find the location (x_{ref}, y_{ref}) of the persons head at the start of such a segment, a sequential Monte Carlo particle filter [8] is used to track the head of the person to get $\{(x^i(t), y^i(t)), w^i(t)\}$ where the superscript denotes the index of the particle and $w^i(t)$ denotes the probability weight for the estimate $(x^i(t), y^i(t))$. Assuming constant velocity models and small motion between successive frames we can show using the optical flow based SfM equations that the angle traced by centroid of the persons head α is related to θ by:

$$cot(\alpha)(x_{ref}, y_{ref}) = \frac{x_{ref} - fcot(\theta)}{y_{ref}}. \qquad (1)$$

Thus, given f and (x_0, y_0), we can compute θ. Knowing (x_0, y_0),$cot(\alpha)$ and θ, f can be computed as part of a calibration procedure.

Having obtained the angle θ, we need to synthesize the canonical view. Let Z denote the distance of the object from the image plane. If the dimensions of the object are small compared to Z, then the variation in θ, $d\theta \approx 0$. This essentially corresponds to assuming a planar approximation to the object. Let $[X_\theta, Y_\theta, Z_\theta]'$ denote the coordinates of any point on the person (as shown in the Figure 1) who is walking at an angle $\theta \geq 0$ to the plane passing through the starting point $[X_{ref} Y_{ref} Z_{ref}]'$ and parallel to the image plane which we shall refer to, hereafter, as the canonical plane. Computing the 3-D coordinates of the synthesized point involve a rotation about the line passing through the starting point and taking the perspective projection we can show that

$$\begin{aligned} x_0 &= f\frac{x_\theta cos(\theta) + x_{ref}(1 - cos(\theta))}{-sin(\theta)(x_\theta + x_{ref}) + f} \\ y_0 &= f\frac{y_\theta}{-sin(\theta)(x_\theta + x_{ref}) + f}, \end{aligned} \qquad (2)$$

where

$$x = f\frac{X}{z} \text{and } y = f\frac{Y}{z}.$$

Equation (2) is attractive since it does not involve the 3D depth; rather it is a direct transformation of the 2D image plane coordinates in the non-canonical view to get the image plane coordinates in the canonical one. Thus knowing the azimuth angle θ we can obtain a synthetic canonical view using (2). Novel view synthesis as described above corrects for the distortion of appearance based features including height and leg-swing [5].

Given $\{\theta^i(t), w^i(t)\}$ derived using the SIS tracker to obtain α and (1), we use the MAP estimate of θ for synthesizing the corresponding probe sequence. Keeping in view the limited training data available and to deal with different number of frames in the gallery and probe a template matching technique based on dynamic time warping [9] is used for recognition. In order to assess the utility of our method without being affected by the choice of a particular image feature, we choose to use the entire image as the feature with binary correlation as a local distance measure. The gait recognition algorithm yields a score which is the cumulative binary correlation distance between the probe video sequence and the gallery.

3. FACE RECOGNITION FROM VIDEO

Face recognition has been an active area of research for a long time. Most of the work in face recognition has focussed on still-to-still situation. Often the probe consists of a video sequence of the unknown subject while the gallery contains static images of the subjects. Different strategies have been developed for this still-to-video scenario. Most approaches involve detecting the face and tracking it over time and when the frame becomes large enough, recognition is performed using still-to-still recognition approaches. Zhou et al. [6] argue against such a tracking-then-recognition since it has unresolved issues like criteria for selecting good frames and estimation of parameters for registration and present a tracking-and-recognition approach which attempts to resolve uncertainties in tracking and recognition simultaneously in a unified probabilistic framework. A time series model is used to fuse temporal information in a probe video, which simultaneously characterizes the kinematics and identity using a motion vector and an identity variable. The joint posterior density of the motion vector and the identity variable is estimated at every time instant and then propagated to the next time instant. Marginalization over the motion vector yields a robust estimate of the posterior distribution of the identity variable. A computationally efficient SIS algorithm is used to estimate the posterior distribution. It was shown that a degeneracy in the posterior probability of the identity variable occurs, leading to improved recognition. The algorithm yields the score in terms of a posterior probability $P(i|X)$ where i denotes a gallery person and X denotes the probe video.

4. FUSION

The ultimate goal of designing pattern recognition systems is to achieve the best possible classification performance for the task at hand. As discussed in [13], fusion of multiple sources of evidence is likely to yield tangible benefits in terms of improved efficiency and accuracy of the identification system. Two different approaches exist for fusion. The first employs multiple experts to provide opinions of the same biometric data e.g. [10] for speaker identification. The second involves using different modalities from the input e.g. [11]. We focus on this approach for fusion.

To improve efficiency of a multimodal biometric system, one can adopt multistage combination rules whereby subjects may be coarsely classified by a less accurate classifier, passing a smaller set of likely candidates to a more accurate classifier. The results of the gait classifier, for example, can be used to pass a smaller

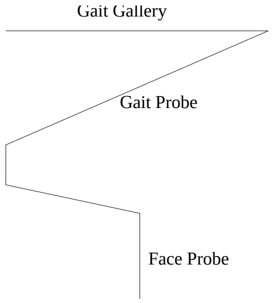


Fig. 2. Inverted Σ pattern used in the NIST database

number of candidates to the more accurate face recognition unit. Alternatively, decisions from the different classifiers can be combined directly using simple rules like SUM, PRODUCT etc.. In this case it is first necessary to transform the scores obtained from the different classifiers in order to make them comparable. The transformation should be such that the relative ordering of the scores is not altered. In other words the transformation function should be monotone. Some of the commonly used transformations include linear, logarithmic, exponential and logistic. The purpose of these transformations is, first, to map the scores to the same range of values and second, to change the distribution of the scores. For example, the logarithmic transformation puts strong emphasis on the top ranks, whereas the lower ranked scores which are transformed to very high values, have a quickly decreasing influence. A detailed discussion of score transformation is given in [12] in the context of combining classifiers for face recognition. The face recognition algorithm yields a match score which is a probability while the gait recognition algorithm yields a distance measurement. In order to make the scores comparable before fusing them, we apply an appropriate transformation to the gait scores. Note that the score transformation is necessary only when the scores of the face and gait recognition algorithms are to be directly combined. In our experiments we describe the results of both the above fusion strategies for face and gait cues.

5. EXPERIMENTAL RESULTS

The NIST database consists of 30 people walking along an inverted Σ -shaped walking pattern as shown in Figure 2. There are two cameras looking at the top horizontal part of the sigma. The camera that is located further away is used in our experiments since the planar approximation we use is more valid in that case. The segment of the sigma next to the top horizontal part is used as a probe. This segment is at an angle 33^0 to the horizontal part. As explained in Section 2, the person's head is tracked using a sequential Monte Carlo filter. Using (1), $\theta \sim (\theta^i(t), w^i(t))$ is obtained. Using $\tilde{\theta}(t) = \arg\max_{w^i}(\theta^i(t), w^i(t))$ the image of the unknown person $X(t)$ is synthesized using Equations 2. A few images from the NIST database are shown in Figure 3. The gait recognition result is shown in Figure 4(a) and (d). The last part of the sequence where the person presents a front view to the camera was used for still-to-video face recognition using [6] and the result is shown in Figure 4 (b) and (e).

We now present the results of the fusion of face and gait cues. As mentioned before, in order to combine the scores from the face and gait classifiers directly, it is necessary to make them comparable. We used the exponential transformation for converting the scores obtained from the gait recognition viz. given that the match score for a probe X from the gallery gaits is given by $S_{X1}, \cdots, S_{XN}$ we obtain the transformed scores $exp(-S_{X1}), \cdots, exp(-S_{XN})$. Finally we normalize the transformed scores to sum to unity. We also tried logistic and logarithmic score transformation methods. The results obtained using these were comparable to the exponential case.

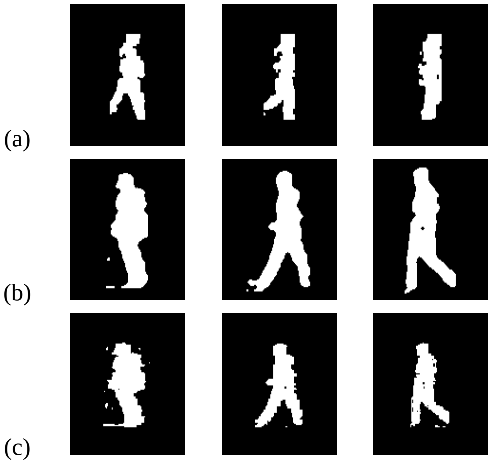


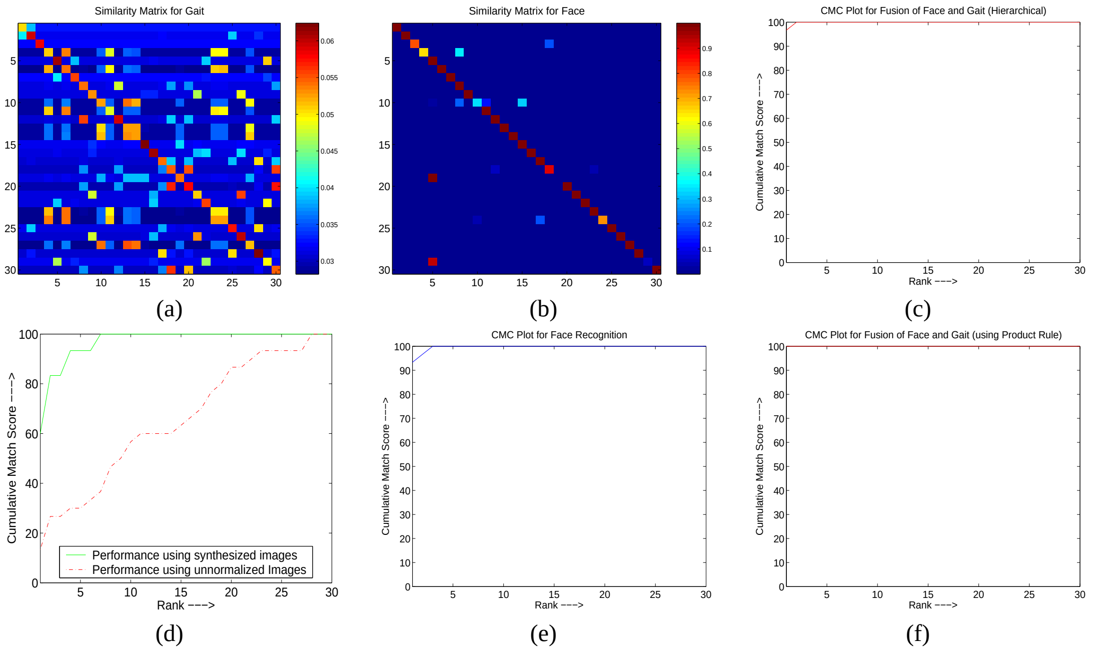
Fig. 3. Examples for the NIST database (a) Gallery images of person walking parallel to the camera (b) Unnormalized images of person walking at 33^0 to the camera(c) Synthesized images for (b).

Hierarchical Fusion: Given the similarity matrix for the gait recognition algorithm, we plot the histograms of the diagonal and non diagonal terms of the normalized similarity matrix. From Figure 5 we note that the distributions of the true matches and false matches have limited overlap. This suggests that a threshold can be determined from the histogram and only individuals whose score is higher than this threshold need be passed to the face recognition algorithm. Although it is tempting to choose this threshold as high as possible, it should be noted that due to overlap in the two histograms, choosing a very high value may lead to the true person not being in the set of individuals passed to the face recognition algorithm . For the NIST database we chose a threshold of 0.035. This results in passing approximately the top six matches from the gait recognition unit to the face recognition algorithm. The CMC plot for the resulting hierarchical fusion is shown in Figure 4 (c). Note that the top match performance has gone up to 97% from 93 % for this case. The more important gain however is in terms of the number of computations required. This number drops to one-fifth of its previous value. This demonstrates the value of gait as a filter.

Holistic Fusion If the requirement from the fusion is that of accuracy as against computational speed, alternate fusion strategies can be employed. Assuming that gait and face can be considered to be independent cues, a simple way of combining the scores is to use the SUM or PRODUCT rule [13]. Both the strategies were tried. The CMC curve for either case is as shown in Figure 4(f). In both cases the recognition rate is 100 %.

6. CONCLUSION

In this paper, we presented experimental results for fusion of face and gait for the single camera case. We considered the NIST database



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4. Similarity matrices for (a) Gait Recognition ; (b)Face Recognition; CMC characteristics for: (d) Gait (e) Face CMC curves for (c) Hierarchical and (f) Holistic Fusion

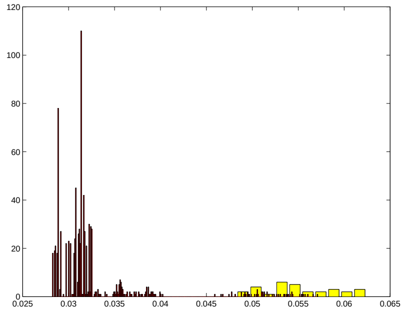


Fig. 5. Histogram of the true matches (yellow) and false matches (black)

which contains outdoor face and gait data for 30 subjects. We used the method described in [5] for gait recognition and the method described in [6] for face recognition. Decision fusion which is a special case of data fusion was to combine the results of the face and gait recognition algorithms. We demonstrated the use of gait as a filter in building more efficient multimodal biometric systems. We also showed the results of directly combining the scores obtained by the individual algorithms to improve the overall recognition rates.

7. REFERENCES

[1] A.K.R Chowdhury and R. Chellappa, "Face reconstruction from video using uncertainty analysis and a generic model," *CVIU*, vol. 91, no. 1-2, July-August.

[2] G.Shakhnarovich, L.Lee, and T.Darrell, "Integrated face and gait recognition from multiple views," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

[3] M.Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[4] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," *Proceedings of the IEEE International Conference on Face and Gesture Recognition*, 2002.

[5] A. Kale, A.K.R Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 143–150, 2003.

[6] Shaohua Zhou and R. Chellappa, "Probabilistic human recognition from video," *Proceedings of ECCV*, 2002.

[7] M.M. Kokar and J.A. Tomasik, "Data vs. decision fusion in the category theory framework," *FUSION 2001*, 2001.

[8] Michael Isard and Andrew Blake, "Contour tracking by stochastic propagation of conditional density," *Proceedings of ECCV*, , no. 1, pp. 343–356, 1996.

[9] A. Kale, N. Cuntoor, A.N. Rajagopalan, B. Yegnanarayana, and R. Chellappa, "Gait analysis for human identification," *Proceedings of 3rd International Conference on Audio and Video Based Person Authentication*, June 2003.

[10] D. Genoud, G. Gravier, F. Bimbot, and G. Chollet, "Combining methods to improve phone based speaker verification decision," *Proceedings of ICSLP*, vol. 3, pp. 1756–1760, 1996.

[11] L. Hong and A. Jain, *Multimodal Biometrics*, Kluwer, 1999.

[12] B. Achermann and H. Bunke, “Combination of classifiers on the decision level for face recognition,” Tech. Rep., Institut fur Informatik und angewandte, Mathematik,, Universitat Bern, 1996.

[13] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, “On combining classifiers,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 226–239, March 1998.