Identifying People based on their Motion Signature

George Williams, Graham W. Taylor, Chris Bregler
Courant Institute of Mathematical Sciences, New York University
{gwilliam,gwtaylor,bregler}@cs.nyu.edu

Introduction

In the past, using motion for recognition has been demonstrated mainly for classifying certain actions or gait styles ([1, 2, 3, 4, 5]). Using the more ambiguous motion of a person that appears during talking and performing "body language" is harder to quantify. This is similar to what the acoustic speech community calls "speaker recognition". It is not important what the person says, but how it is said. We developed a new video-based feature extraction technique and methods to train statistical models that classify body motion signatures. The recognition architecture is inspired by recent progress in speaker recognition research. This abstract has 3 main contributions: 1) a visual feature estimation technique, based on sparse flow computations and motion angle histograms that we call "Motion Orientation Signatures" (MOS). 2) Integration of this feature into a 3-stage recognition system, 3) an integration method with a state-of-the-art face recognition architecture [6]. We demonstrate how this new technique can be used to identify people just based on their motion, or it can be used to significantly improve "hard-biometrics" techniques. For example, face verification achieves on this domain 6.45% Equal Error Rate (EER), and the combined verification performance of motion features and face reduces the error to 4.96% using an adaptive score-level integration method. The more ambiguous motion-only performance is 17.1% EER. This is to the best of our knowledge the first time a system has been demonstrated that can significantly improve a state-of-the-art face recognition system with such complex and ambiguous signals like a person's body-language.

Motion Orientation Signatures (MOS)

The first step in our new visual extraction schema is flow computation at reliable feature locations. Given these robust flow estimates, we compute weighted flow angle histograms. (See [7] for more details on the visual feature extraction method). Inspired by acoustic speech features, we also compute "delta-features", the temporal derivative of each orientation bin value. Since the bin values are statistics of the visual velocity (flow), the delta-features cover acceleration and deceleration. Figure 1 shows a few examples that demonstrate what signatures are created with certain video input. We show several visualizations of these features at http://movement.nyu.edu/GreenDot

We further improved our MOS feature extraction theme in incorporating a state-of-the-art face-tracking system from PittPatt.com [6] (winner of the 2008 NIST MBGC Challenge), and placing a $N \times M$ grid of local region of interests (ROIs) around the face such that it covers the body. Inside each local ROI of the grid we compute the MOS feature and concatenate all local ROI histograms to one big feature vector.

GMM-Super-Features: Similar to recent approaches in the speech community we convert an arbitrary length video into a fixed dimensional feature vector with so called Super-Features [8]: A Gaussian Mixture Model is first trained on the MOS features. In speaker recognition, this is called the Universal Background Model (UBM). Given that UBM model, the statistics of each video are computed in MAP adapting the GMM to the specific video features. A GMM-Super-Feature is the difference between the UBM mean vectors and the new MAP adapted mean vectors. If the new video has some unique motion, then at least one mean vector has a large difference to the UBM model.

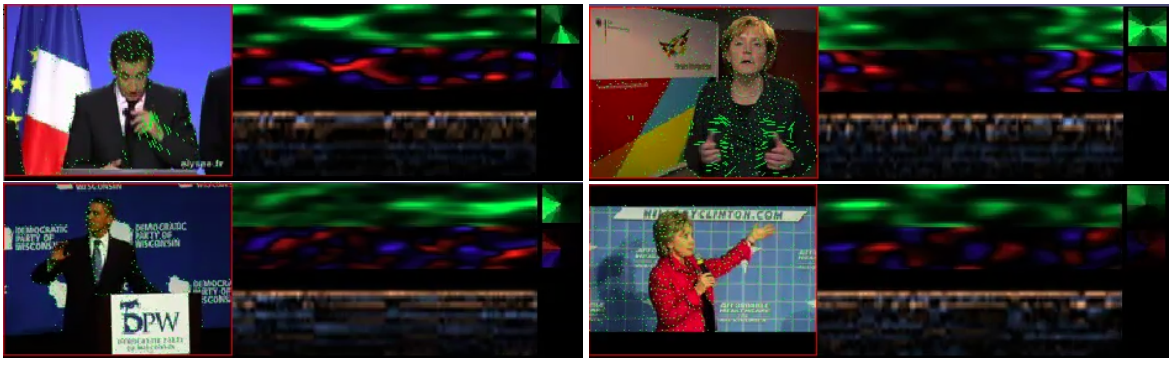

Figure 1: Examples of Motion Signatures. The green colored signature shows how the orientation of the sparse flow features changes over time (top row), and the red-blue color coded signature shows the delta features (middle row).

Classification: In [7] we feed those GMM-Super-Features into a standard SVM classifier. However we have subsequently found that a logistic regression classifier applied to the GMM-Super-Features achieves the same performance. All further results reflect logistic regression.

Comparison with Bag-Of-Feature Architectures: Our approach somewhat is related to so called Bag-Of-Feature architectures used by [1, 2, 3, 4] for activity analysis, but our experiments show that: 1) When we use so called Space-Time Interest Points (STIP) [1] instead of our MOS features, we get an relative increase in error 17.1% to 25%. 2) We also replaced the GMM-Super-Features with higher-level histograms of vector-quantized codewords (visual words), as it is used in all Bag-Of-Feature architectures, and our relative error rate increased further to 31% EER. This is not surprising, since during the vector quantization much information is lost, that is retained by the GMM-Super-Features.

Experiments

In total we have 72 minutes of video data with 320×240 pixel resolution. For each subject we have at least 4 different video clips recorded at different times. Most subjects are public figures like international politicians or talk show hosts. We are currently also collecting a database of "non-famous" people.

We trained our system to verify if a specific person is present in the video. For positive examples we chose two or more videos of that subject, and for negative examples we sampled from the training set part of the remaining videos in our database. We repeated each experiment 27 times (3 different random initializations for the UBM model, and 9 different verification targets). Figure 2 shows the ROC curve of our performance. In the speaker verification community it is common to report the Equal Error Rate EER (the error at the diagonal of the ROC curve, when the number of false positives and false negatives is equal). For this case we achieved 17.1%.

The crucial experiment now is to test if such a soft-biometric modality can be applied in a multi-modal system consisting of other hard-biometrics. We trained a state-of-the-art face recognition system on the same video database. Usually face recognition requires high-resolution input images that have at least 90 pixels between the eyes. In our case the eye distance can be as small as 20 pixels. PittPatt [6] achieved previously the best face recognition results on video recordings at the NIST MBGC 2008 Challenge. We trained the PittPatt system on our low-resolution database, and achieved verification performance of 6.45% EER (figure 2).

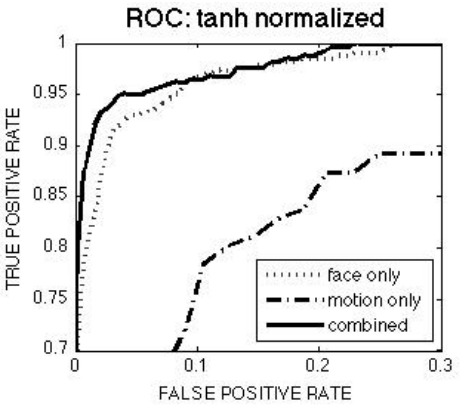

Figure 2: ROC curves using score intergration techniques.

Previously we experimented with multi-modal integration on the input-level [7]. In this abstract we focus on output score-level integration, which allows us to combine several systems that are trained with differently (like PittPatt's face system, and our motion based system). There are many ways to combine scores [9]. We experimented with $\tanh$ normalization [9] and training another logistic regression on the face and motion scores to predict the multimodal score. We achieved best results with the $\tanh$ score combination which reduced the error relatively by 23% to 4.96% EER. The logistic regression achieved slightly less performance, but still a significant improvement over each modality alone to a combined error of 5.24 %.

Conclusion

We have demonstrated how the body motion of a subject during speaking and gesturing can be measured and analyzed with a new visual feature extraction technique and methods that have been successfully applied in the speaker recognition community. Most importantly, we have shown how we can significantly improve the performance of a hard-biometrics system, specifically a state-of-the-art face recognition system with a new soft-biometrics roughly characterized as "body language". This hints strongly that the anecdotal evidence is true: everybody moves in a unique individual way, and we can use this feature for person identification.

References

[1] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *Proc. CVPR*, 2008.
[2] J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.
[3] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.
[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *ICCV VS-PETS*, 2005.
[5] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer. The humanID gait challenge problem: data sets, performance, and analysis. *IEEE PAMI*, pages 162–177, 2005.
[6] PittPatt.com. PittPatt Face Detection, Tracking, and Recognition SDK.
[7] C. Bregler, G. Williams, S. Rosenthal, and McDowall I. Improving Acoustic Speaker Verification with Visual Body-Language Features. In *Proc. ICASSP*, 2009.
[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13:308–311, 2006.
[9] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285, 2005.

Topic: visual processing and pattern recognition Preference: oral or poster