RESEARCH BRIEF

Institute for Homeland Security Solutions
Applied research • Focused results

June 2010

Human Identification from Video: A Summary of Multimodal Approaches

Project Leads

Charles Schmitt, PhD, Renaissance Computing Institute

Allan Porterfield, PhD, Renaissance Computing Institute

Sean Maher, MS, MBA, Cambridge Intelligent Systems

David Knowles, MA, Renaissance Computing Institute

Statement of Problem

The U.S. Department of Homeland Security's (DHS) Personal Identification Systems Thrust Area focuses on developing an "accurate, contactless, near real-time capability to identify known threats…through effective, interoperable multi-biometrics capabilities" (DHS Web site, 2009). A robust ability to accurately identify people from video under a variety of real-world conditions is an important capability in this context because it is potentially one of the most accurate and widely deployable technologies that requires neither contact nor consent.

Face recognition from still images is one of the leading image-based biometrics. In applications where illumination and view angle are highly controlled and image resolution is high, recognition rates for the leading still image systems is very high. The best performing software in the 2006 Face Recognition Vendor Test (FRVT) achieved a false accept rate (FAR) of 0.01 at a false reject rate of 0.001 (Phillips et al., 2007).

Despite the high performance of still image systems in controlled conditions, performance can drop quite rapidly as variation in view angle, illumination, occlusion, or viewing distance increases, or as image resolution decreases. Heo, Abidi, Paik, and Abidi (2003) showed that identification rates for FaceIt, the best performing software from the 2002 FRVT, and one of the best performers in the 2006 FRVT, dropped below 60% when illumination or view angle was varied, and dropped as low as 0% when viewing distance was increased from 3 ft. to between 9 ft. and 12 ft. Arandjelović and Cipolla (2006) also ran experiments with large variations in view angle and illumination and found that FaceIt achieved an identification rate of only 64%. Recent research by Singh, Vatsa, and Noore (2009) also indicate that plastic surgery can seriously degrade recognition rates. In several experiments using pre- and post-surgery images, the authors found post-surgery recognition rates of 10.6% to 38.8%, depending on the extent of surgery.

Maintaining high verification and identification rates when there is significant variation in illumination, view angle, degree of occlusion (eyeglasses, scarves, beards, etc.), surgical alteration, or viewing distance continues to present challenges to building truly robust person-identification systems that operate well in less controlled imaging conditions. This brief will investigate the advantages of using multimodal systems to improve accuracy and robustness of video-based biometric identification systems.

Background

Many approaches, including automatic face recognition, speaker identification from audio channels, ear image recognition, and gait analysis, have been developed to automatically identify people from video. More recently, researchers have looked at multimodal approaches that augment face recognition with additional signal sources and algorithms to improve identification accuracy and robustness. This brief will summarize current research combining face and ear images, visible spectrum and thermal spectrum face images, face and speech signals, and face and gait information.

The analyses below are meant to summarize the improvement in identification accuracy and robustness attributable to combining modalities. Because researchers often use different video databases to test their algorithms, comparison of raw performance metrics across studies would not be meaningful without a much more detailed analysis that is beyond the scope of this brief.

Face and Ear Images

The ear, like a fingerprint or iris, is an information-rich anatomical feature that can be used to identify individuals. Although less researched than other biometric features, technologies that recognize the ear are receiving increasing interest and seem to show particular promise when combined with face recognition.

Chang, Bowyer, and Barnabus (2003) found that combining face and ear images improved recognition rates by over 20% relative to either modality alone. In one representative experiment, both their face and ear recognition algorithms had recognition rates of around 71%, while the combined system achieved 91%. Theoharis, Passalis, Toderici, and Kakadiaris (2008) achieved a 99.7% recognition rate through the fusion of face and ear images whereas face recognition alone identified 97.5% correctly, and ear recognition alone achieved a 95.0% recognition rate. Yan (2006) achieves a 100% recognition rate with a combined face and ear recognition system. The recognition rate for face alone was 93%, and ear alone was 98%.

Combining face and ear images provides robustness, primarily with respect to view angle. As the view angle deviates from a full-frontal view, most face recognition algorithms degrade in performance (Heo et al., 2003). Simultaneously a better image of the ear is available. The effect of occlusions such as hair, hats, and earrings on the performance of ear recognition algorithms is unclear, as is the robustness of current ear recognition algorithms to angle of view and illumination changes.

Visible and Thermal Spectrum Face Images

Using thermal imagery for person identification has two very distinct characteristics. First, thermal imagery, especially in the longwave infrared (LWIR) spectrum, is invariant to changes in illumination, since it measures heat emitted by a person's skin, not reflected light from external sources. Second, most facial disguise techniques alter the thermal signature of the face enough that disguises are detectable with infrared (IR) cameras (automatic identification is still difficult or impossible, but the presence of a disguise is detectable). IR spectrum images can even be used to detect surgical alterations, since even subcutaneous scars appear as distinct cold spots on thermal imagery (Kong, Heo, Abidi, Paik, & Abidi, 2005).

Arandjelović, Hammoud, and Cipolla (2006) achieved a 97% recognition rate using a combination of visible- and thermal-spectrum face images on a dataset whereas their visible-spectrum face recognition algorithm achieved 87% by itself, and their thermal-spectrum algorithms achieved 82%. Gundimada and Asari (2009) evaluated several algorithms and fusion techniques. Their best approach achieved a 99.24% recognition rate whereas the visible and thermal algorithms achieved 98.11% and 95.50% respectively. Chen, Flynn, and Bowyer (2003) showed that using fairly simple algorithms based on principal components (PCA), they were able to achieve a 91% recognition rate by combining visible and thermal imagery on a database where one of the best commercial systems (FaceIt) achieved 86% using visible images only. In their experiments, both the visible only and thermal only PCA algorithms achieved recognition rates around 75%. Heo, Kong, Abidi, and Abidi (2004) used FaceIt as an individual module and achieved 93% accuracy for both visible and thermal images (independently) and 98% when combined. Scolinsky and Selinger (2004) experimented with images captured indoors as well as outdoors. For indoor images, their system achieved 98.4% accuracy whereas the individual visible and thermal modules achieved 97% and 94%

respectively. For outdoor images, their system achieved 89% accuracy whereas the individual visible and thermal modules achieved 67% and 83% respectively. The latter experiment clearly shows both the degraded performance due to illumination variance in outdoor images as well as the superior performance of IR image–based algorithms when there is significant variation in illumination.

Face and Speech Recognition

Since most video cameras provide both video and audio channels, several researchers have developed ways of combining face-recognition and speaker-identification algorithms to improve the identification accuracy relative to either independent modality. Both text-dependent and text-independent speaker identification algorithms are used. With the former, the same phrase must be spoken in the enrollment and identification stages whereas the latter attempts to identify the speaker regardless of word choice. Generally speaking, text-dependent systems perform better but have the constraint of requiring a cooperative subject.

By combining face and text-dependent speaker identification in a mobile identification system, Hazen, Weinstein, and Park (2003) achieved a 50% reduction in equal-error rate (EER) in a user verification system relative to either independent modality. In their experiments, the combined face-speech method had an overall accuracy of 98%. Ben-Yacoub, Abdeljaoued, and Mayoraz (1999) achieved a 70% reduction in equal-error rate and an overall accuracy of 96% by combining face recognition and text-independent speaker identification. Brunelli and Falavigna (1995) achieved 98% accuracy with a combined system whereas face recognition alone achieved 91% and speaker identification alone achieved 88%.

The greater robustness from combining face recognition and speaker identification derives mainly from the independence of the signal sources. That is, in some circumstances (because of illumination, view angle, etc.) face recognition may be very difficult, but a clear audio signal may be present. Likewise, there may be situations where the audio environment is very noisy, but the visual environment is clean.

Face and Gait Recognition

Most biometric techniques such as face recognition, speaker identification, and ear recognition (as well as iris and fingerprint biometrics) require the subject to be fairly close to the sensors used for data acquisition. Gait biometrics attempt to create a unique signature for a person based on the way the person walks and thus offer the potential of identifying subjects at a greater distance than other techniques.

Kale, Roy-Chowdhury, and Chellappa (2004) evaluated fusion of frontal-face and gait-recognition algorithms on a National Institute for Standards and Technology (NIST) database, where their face recognition algorithm identified 97% of subjects correctly, and the gait recognition alone identified 60% correctly. When combined, they achieved perfect identification. In a set of somewhat more challenging experiments, Zhou and Bhanu (2007)

combined profile face recognition with gait identification to achieve a 91% combined accuracy whereas profile face recognition alone achieved 80%, and gait recognition achieved 82%. In a recent paper, Geng, Wang, Li, Wu, and Smith-Miles (2009) achieved an 86% recognition rate by adaptively combining face and gait subsystem classifiers based on analysis of the view angle and subject-to-camera distance on a rather challenging dataset that included large variations in these two dimensions.

Synthesis

Notwithstanding the high recognition rate in highly controlled conditions, performance of current state-of-the-art image- and video-based biometric systems degrades significantly when presented with real-world variation in illumination, view angle, degree of occlusion, and viewing distance. Combining modalities can improve robustness with respect to these variations and achieve higher verification and identification rates than single-modality systems. There is clear evidence that significant progress continues to be made in handling real-world variations as current state-of-the-art algorithms in the research literature significantly outperform existing commercial software. This is clearly seen, for example, in Arandjelović and Cipolla (2006) where they achieve 99.7% accuracy on a dataset with extreme illumination and pose variation. On the same dataset, FaceIt achieved only 64% accuracy.

Future Directions

Operational Relevance

Perhaps the single largest gap in the research community is analyses of how systems perform in specific operational scenarios of interest to DHS. Better understanding the variations in view angle, illumination, occlusion, viewing distance, and possibly other parameters actually encountered in practice at air, land, and sea entry and exit points would be extremely helpful in characterizing the likely performance of multimodal biometric systems in real-world situations. Further, better understanding the most relevant performance metrics for various operational scenarios would be very helpful, as there is often not a simple mapping between metrics used in the literature (FAR, EER, rank-1 identification rate, etc.) and real-world utility.

Improved Performance in Operational Environments

Based on the research reviewed, it is clear that while standoff, video-based, multimodal systems reviewed here often perform significantly better than their unimodal subsystems, the absolute performance levels for many operational environments are likely too low to provide significant value to DHS. Designing systems to perform well in specific operational environments would be very valuable.

Standard Databases

A set of standard databases designed to characterize the performance of video-based biometrics with respect to variations in view angle, illumination, occlusion, viewing distance, and other parameters identified in specific operational scenarios would be extremely valuable and do not currently exist. Other variations such as facial expression, ethnicity, reference database cardinality, and various disguise techniques could be included as well.

More Intelligent Fusion

To date, only pairs of modalities have been researched. Because each modality has specific strengths (e.g., thermal imagery provides robustness with respect to illumination variance, while gait recognition provides viewing distance robustness), combining more modalities likely will further improve both accuracy and robustness. In addition to combining more modalities, current state-of-the-art research could be extended by creating fusion algorithms that more intelligently combine modalities. Work in this direction has been started by Arandjelović and Cipolla (2006), who show that automatically detecting the difficulty of the "illumination problem" in each instance before deciding how to combine visible and thermal subsystem classifications improves accuracy, as well as by Geng et al. (2009), who show that adaptively combining face and gait subsystem classifiers based on analysis of the view angle and subject-to-camera distance significantly improves performance over static combination rules. The impact of such fusion on identification-verification performance and deployment cost is an additional unaddressed area of research.

Computational Performance

Video-based biometric techniques are computationally resource-intensive. For example, one the best performing video-based face recognition algorithms has time complexity of roughly $O(n^3)$, where n is the number of frames (Arandjelović, personal communication, 2009). Although more accurate and robust, multimodal techniques demand even more computing resources, given that multiple computationally intensive algorithms must run simultaneously. There has been a dearth of research in understanding the time and space complexity of various algorithms and in developing optimized software and hardware for handling the most computationally intensive sub-algorithms. In particular, identifying bottleneck steps in the highest performing algorithms and developing software and hardware systems to optimize the performance of those steps would be very beneficial in translating state-of-the-art research into usable systems. Similarly, developing metrics that combine identification-verification performance and computational complexity to allow more transparent analysis of the cost associated with improved identification-verification performance would be quite valuable in evaluating multimodal biometric systems.

Contact Information

Dr. Charles Schmitt

Renaissance Computing Institute, Suite 540, 100 Europa Dr.
Chapel Hill, NC 27517

919-445-9696

cschmitt@renci.org

Dr. Allan Porterfield

Renaissance Computing Institute, Suite 540, 100 Europa Dr.
Chapel Hill, NC 27517

919-445-9611

akp@renci.org

Sean Maher

305 Columbia Place East
Chapel Hill, NC 27516

215-932-2103

sean.maher@cantab.net

David Knowles

Renaissance Computing Institute, Suite 540, 100 Europa Dr.
Chapel Hill, NC 27517

919-445-9677

dknowles@renci.org

Charles Schmitt, PhD, is a senior researcher in data mining and the manager of health informatics for the Renaissance Computing Institute. Dr. Schmitt has a PhD in computer science from the University of North Carolina at Chapel Hill, where he researched the application of neural networks to the field of image analysis and pattern recognition, with a focus on understanding biological visual systems. After graduate school, Dr. Schmitt worked in industry in the field of data mining and software engineering, most recently working in the domain of bioinformatics.

Allan Porterfield, PhD, received his PhD from Rice University in computer science, studying compiler optimization of cache usage. He spent 17 years at Tera (later named Cray, Inc.) as part of a small group (two to five members) responsible for design and implementation of all language-related tools, including a highly optimizing parallel compiler, runtime libraries, incremental linker, debugger, performance tools, and instruction set simulator. He moved to

the Renaissance Computing Institute in 2006 and joined the High Performance Computing and Performance Tool Group looking at ways to increase application performance on large systems. He is principal investigator of the MAESTRO project, funded by the U.S. Department of Defense, looking into runtime designs that maximize the capabilities of modern computer systems.

Sean Maher, MBA, is president of Cambridge Intelligent Systems (CIS). CIS conducts advanced research and development in intelligent video processing and machine-learning algorithms. Mr. Maher has done graduate work in neuroscience, psychology, and computer science and holds an MBA from the University of Cambridge.

David Knowles, MA, is the director of economic development and regional engagement at the Renaissance Computing Institute (RENCI). He manages RENCI's economic development initiatives, entrepreneurship programs, and campus-based engagement centers throughout the state. He comes to RENCI from Georgia Tech, where he was business development manager for the Southeastern Trade Adjustment Assistance Center, a division of the university's Economic Development Institute. Previously, Mr. Knowles was vice president of operations for Interra International, Inc., a leading global food trading company, and chief operating officer of International Trade Management, Inc., an Atlanta-area transactional software firm. Knowles studied international relations at the University of Aberdeen in Scotland and also holds a BA in political science from the University of Missouri and an MA in international relations from the University of Virginia.

References

Arandjelović, O., Hammoud, R., & Cipolla, R. (2006). Towards person authentication by fusing visual and thermal face biometrics. *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance,* 50–56.

Arandjelović, O., & Cipolla, R. (2006). Face recognition from video using the generic shape-illumination manifold. *Proceedings of the IEEE European Conference on Computer Vision*, *4*, 27–40.

Ben-Yacoub, S., Abdeljaoued, Y., & Mayoraz, E. (1999). Fusion of face and speech data for person identity verification. *IEEE Transactions on Neural Networks, 10*(5), 1065–1074.

Brunelli, R., & Falavigna, D. (1995). Personal identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17*(10), 955–966.

Chang, K., Bowyer, K., & Barnabas, V. (2003). Comparison and combination of ear and face images in appearance based biometrics. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 25*, 1160–1165.

Chen, X., Flynn, P. J., & Bowyer, K. W. (2003). Visible-light and infrared face recognition. *Computer Vision and Image Understanding, 99*(3), 332-358.

Geng, X., Wang, L., Li, M., Wu, Q., & Smith-Miles, K. (2009). Adaptive fusion of gait and face for human identification in video. *Proceedings of the 2008 IEEE Workshop on Applications of Computer Vision, 0*, 1–6.

Gundimada, S., & Asari, V. K. (2009). Facial recognition using multisensor images based on localized kernel eigen spaces. *IEEE Transactions on Image Processing, 18*(6), 1314–1325.

Hazen, T. J., Wienstein, E., & Park, A. (2003). Towards robust person recognition on handheld devices using face and speaker identification technologies. *Proceedings of the 5th International Conference on Multimodal Interfaces,* 289–292.

Heo, J., Abidi, B., Paik, J., & Abidi, M. A. (2003). Face recognition: Evaluation report for FaceIt®. *Proceedings of the SPIE 6th International Conference on Quality Control by Artificial Vision (QCAV03)*, *5132*, 551–558.

Heo, J., Kong, S. G., Abidi, B. R., & Abidi, M. A. (2004). Fusion of visual and thermal signatures with eyeglass removal for robust face recognition. *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04), 8*, 122.

Kale, A., Roy-Chowdhury, A. K., & Chellappa, R. (2004). Fusion of gait and face for human identification. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04),* 5, 901–904.

Kong, S. G., Heo, J., Abidi, B. R., Paik, J., & Abidi, M. A. (2005). Recent advances in visual and infrared face recognition. *Computer Vision and Image Understanding, 97*, 103–135.

Phillips, J. P., Scruggs, T. W., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., & Schott, C. L. (2007). *FRVT 2006 and ICE 2006, large-scale results (NISTIR 7408)*. Gaithersburg, MD: National Institute of Standards and Technology

Scolinsky, D., & Selinger, A. (2004). Thermal face recognition in an operational scenario. *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04), 2*, 1012–1019.

Theoharis, T., Passalis, G., Toderici, G., & Kakadiaris, I. A. (2008). Unified 3D face and ear recognition using wavelets on geometry images. *Pattern Recognition, 41*, 3, 796–804.

Yan, P. (2006). *Ear biometrics in human identification*. PhD thesis, University of Notre Dame.

Zhou, X. L., & Bhanu, B. (2007). Integrating face and gait for human recognition at a distance in video. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics 37*(5), 1119–1137.

U.S. Department of Homeland Security. (2009, August 10). Science & Technology Directorate Human Factors/Behavioral Sciences Division. Retrieved from http://www.dhs.gov/xabout/structure/gc_1224537081868.shtm