https://doi.org/10.2352/EI.2023.35.5.IRIACV-326

Automated Dataset Collection Pipeline for Lip Motion Authentication

Shad Torrie, Andrew Sumsion, Zheng Sun, Dah-Jye Lee

Abstract

Biometric authentication takes on many forms. Some of the more researched forms are voice and facial authentication. Due to the amount of research in these areas, there are benchmark datasets easily accessible for new researchers to evaluate new methods. A newer, less researched biometric method is lip motion authentication. These systems entail a user producing a lip motion password to authenticate, meaning they must utter the correct word or phrase to gain access. Because this method is less researched, there are no large-scale datasets that can be used to compare methods as well as determine the actual levels of security that they provide. We propose an automated dataset collection pipeline that extracts a lip motion authentication dataset from collections of videos. This dataset collection pipeline will enable the collection of large-scale datasets for this problem thus advancing the capability of lip motion authentication systems.

Introduction

In recent years, biometric authentication methods have become fields of immense research, controversy, and innovation. Biometrics have become commonplace in consumer electronics as well as commercial and private security systems. The largest advantage of a biometric-based authentication system is that only the individual can authenticate. While this is the ideal case, often there are ways to spoof the systems. These spoof techniques vary depending on the biometric method. Zhou et al. found that 3D face identification systems can often be fooled by projecting facial images onto a 3D head model [1]. Shmelkin et al. created a Generative Adversarial Network (GAN) that generates what they name "master faces" that are able to authenticate as many different individuals [2]. It is also possible, albeit unlikely, for an individual's biometric to be similar enough to pass on another individual's. This can occur more often when subjects are related. These vulnerabilities are worrisome when biometrics are used to secure very important and sensitive information or locations. The large downside to these forms of biometric authentication is that they are not re-securable. Conventional passwords, although with their own challenges, do not have this issue, if the password is phished or discovered, the user need only reset their password to re-secure the system. Biometric systems are much less re-securable due to their nature, you only have one face and one voice. Adding a second layer to biometric systems that adds the re-securable strength of conventional passwords is invaluable. Lip motion authentication is one such added layer.

Lip motion authentication is a biometric authentication method that is easily paired with facial authentication to create a two-factor authentication system that is re-securable if the integrity is compromised. These systems work by having the subject enroll a lip motion password by uttering (vocally or sub-vocally) a word or phrase. The motion of their face is encoded as a separate authentication token to be compared when a user attempts to authenticate. The lip motion authentication acts as a liveness detection layer thus eliminating the ability to use static images or 3D masks to spoof the system. It also ensures attention and intention meaning that the user must intentionally and attentively act in order to authenticate thus eliminating unintentional authentication or forced authentication. Despite these additional layers of security, as with other biometric systems, there are ways that the system can be spoofed or fooled by an individual not enrolled, but if this does occur, the owner need only reset their lip motion password to restore the integrity of the system.

The prospects of lip motion authentication are exciting and there have been promising results achieved by various research groups as seen in Table 1. The datasets used to yield these results, however, are on small groups of people with low amounts of utterances and very limited demographic variance. It can also be observed that there are many methods to approach the lip motion authentication task. Some papers combine voice authentication with lip motion authentication [1, 3–8] while others combine facial recognition with lip motion authentication [9–14]. How each system analyzes lip motion and what constitutes authentication also varies. Some implementations authenticate by ensuring the user utter the same stored lip motion password and others authenticate by ensuring the user moves their face in the same manner, effectively saving how the person moves their face in general to authenticate. This makes it difficult to compare results from the different approaches as well. Other more proven biometric methods have large highly variant datasets that firstly solidify the use cases and method of authentication and also ensure more real-world generalization. A large-scale, highly variant dataset is required to unify the research in this new form of authentication. It will thus enable a more focused direction for future advancements and improvements.

Background

Due to the nature of neural networks, biometric authentication methods require large-scale datasets to ensure their applicability to the variation in the real world. With more popular authentication methods such as facial recognition, there are immense, popular benchmark datasets that can be used to evaluate new advancements in neural network architectures, training strategies, etc. One such example is the VGGface2 dataset [15] which is commonly used when evaluating new methods of performing facial recognition. It contains over 3.3 million facial images from more than 9,000 subjects.

In the lip motion authentication world, there are few publicly available datasets. This results in most works in this area collecting and evaluating on custom datasets as can be seen in

Table 1. The most common publicly available dataset to use as a benchmark is the XM2VTS dataset [16] which contains only 7,080 videos of 295 subjects. This dataset was collected in 1999 in a controlled environment. It thus contains little to no variation in lighting, head position, and vocabulary. This makes it a very non-ideal dataset for comparing various lip motion authentication methods.

Wright et al. found that training on the VM2VTS dataset resulted in poor performance on varying real-world data particularly when lighting conditions change [34]. They found that their method resulted in a 1.21% error when trained and evaluated on the XM2VTS dataset. They then collected their own more real-world dataset named FAVLIPS on mobile devices in varying lighting conditions. The FAVLIPS dataset contains an additional 2,268 videos of 42 individuals. These videos had four specific lighting conditions and were collected across 4 sessions that occurred 4 weeks apart. They found that training only on XM2VTS results in poor results when lighting conditions vary (as seen in Table 2). They then trained with the FAVLIPS dataset added which resulted in much preferable error rates. While these results are encouraging, the FAVLIPS dataset is still no where close to the large-scale dataset required to validate lip motion authentication methods as a preferable biometric method.

A highly researched area that is very similar to lip motion authentication is automatic lip reading. This entails predicting what a person is uttering with vision only. There are many datasets collected for these systems. Unfortunately, most of these datasets don't distinguish individuals identities thus making it difficult to use them for training and evaluating purposes on lip motion authentication systems. The OuluVS dataset is one exception [21]. It contains 20 specified individuals which is small when training a biometric system to generalize to the population. [10] utilized the OuluVS dataset to train their lip motion authentication methods and found decent results (see table 1). The size of this dataset limits its applicability to the biometric problem.

The largest datasets in the area of automated lip reading are collected by scraping videos of people speaking from TV channels or videos from the internet resulting in datasets with thousands of individuals and hundreds of thousands of videos [36–44]. Due to the large real-world variance in these datasets they are often referred to as "in the wild" datasets. They have brought large advancements and large challenges in automated lip reading research. We propose a pipeline similar to the ones used in the collection of these datasets to enable the collection of in-the-wild datasets for lip motion authentication.

Methods

The automated dataset collection pipeline that we propose in this work will be used in future works to collect large in-the-wild datasets for the lip motion authentication task from public video footage on TV channels or internet streaming services like YouTube. The pipeline contains the following steps (as illustrated in Figure 1): Identify the individual, determine when they are speaking, determine what words are spoken, determine the beginning and end timestamp for each word in the video and save the relevant information in a biometric dataset format for future use.

Firstly, individual identification will be done by performing facial recognition on the individuals throughout the videos in question. We chose the FaceNet facial recognition network

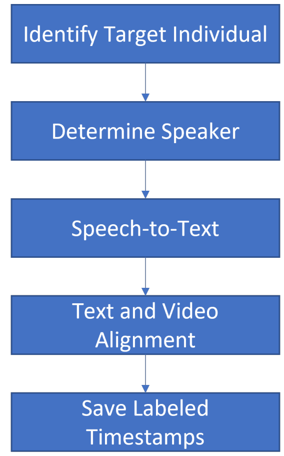

Figure 1: This flowchart illustrates the automated dataset collection pipeline described.

to perform this step [45]. The facial recognition embeddings are then compared to determine where in each frame of each video the subject is.

The next step is to determine when the individual in question is speaking. We utilize the SyncNet toolkit [46] to analyze the audio and video and determine the likelihood that the person is speaking in each frame. This is then compared with the facial recognition data to determine when the individual in question is speaking. The smaller sections of speech by the individual are then removed.

Once the timestamps of when the individual in question is speaking are found, we then must determine what the individual is saying. To do this we use Google's text-to-speech recognition. Unfortunately, this does not give the timestamps for when each word is uttered. Thus, the next step is to align the text with the video and determine the start and end time stamps for each word uttered. This is done with the Penn Phonetics Lab Forced Aligner [47].

Once the beginning and end timestamps are found for each word uttered by the individual the data must be collected in the correct structure. For the lip motion authentication task, the data must be organized to allow for multiple instances of the same individual and word as positive cases, instances of the individual and a different word as negative samples, and videos of other individuals speaking any words as negative samples. Thus data is organized first by the individual, then by the word spoken, and then by the time stamp and video for each instance of the individual speaking the given word.

This pipeline is similar to others made to collect automatic lip-reading datasets, but to our knowledge, this type of pipeline has not been applied to the lip motion authentication task. This adds extra complexity to the processing and labeling of the data as well as the ramifications of incorrect labeling.

Results

To test out the described pipeline we used the same videos used in the VoxCeleb2 dataset [48]. This is a voice recognition dataset that contains about six thousand individual speakers spread across 145K videos extracted from YouTube. They report

Paper(year)	Method(s)	Dataset	Speakers	Metric	Results
[17](2020)	motion	AV Digits [18]	39	EER	9%
[9](2006)	face+motion	BioID [19]	25	Accuracy	86%
[20](2003)	motion	M2VTS	36	EER	19.7%
[10](2018)	face+motion	OuluVS [21]	20	Accuracy	71%**,93.25%
[22](2004)	motion	private	40	EER	5.1%
[23](2011)	structrue+motion	private	21	Accuracy	99.5%
[24](2013)	motion	private	43	FAR@ 3% FRR	14.5%
[25](2014)	motion	private [26]	20	Accuracy	92.4%
[27](2017)	motion	private	20	Accuracy	96.2%
[28](2018)	UT motion	private	50	TNR&TPR	86.7%&76.7%
[1](2021)	audio+voice+face	private	44	EER	5%
[7](2021)	voice+motion	private	240*	FAR&FRR	0.25%&18.25%
[8](2021)	voice+motion	private	50	Accuracy	95.89%
[14](2022)	face+motion	private	48 [13]+11	AP	98.8%**
[11](2018)	face+motion	UvA-NEMO [29] & KAIST [11]	400&104	EER	0.37%
[3](2004)	voice+motion	VidTIMIT [30]	43	EER	1.0%
[31](2000)	motion	XM2VTS	295	EER	14%
[4](2006)	voice+motion	XM2VTS	295	EER	22%**,2%
[5](2007)	voice+motion	XM2VTS	295	Accuracy	78%**,98%
[6](2012)	voice+motion	XM2VTS	295	Accuracy	94.7%
[32](2015)	motion	XM2VTS	295	EER	2.2%
[33](2019)	motion	XM2VTS	295	EER	1.03%
[34](2020)	motion	XM2VTS	295	EER	1.65%
[35](2022)	motion	XM2VTS	295	Accuracy	96.78%

Table 1: Comparison of lip motion authentication systems' datasets used, their size, and the results achieved.
*Number of videos (number of speakers not provided)
**Motion only results

	Training Data		
	XM2VTS only	**pretrained on XM2VTS updated on FAVLIPS**	**XM2VTS+FAVLIPS**
XM2VTS: Evaluation Set	1.21%	1.95%	5.60%
FAVLIPS: Neutral Nums	22.43%	13.79%	10.83%
FAVLIPS: Light Front	28.44%	17.50%	20.54%
FAVLIPS: Light Side	42.29%	36.67%	30.00%
FAVLIPS: Light Behind	44.91%	24.17%	29.12%

Table 2: [34] preformed this evaluation on the XM2VTS dataset [16], and sections of the FAVLIPS [34] dataset with a comparison of which training data was used. Evaluation metric reported in EER (Equal Error Rate).

over 1 million utterances, an utterance being an entire sentence used for voice recognition. They have previously labeled when the individual in question is speaking having utilized SyncNet and Penn Phonetics Lab Forced Aligner as we have in our pipeline.

Running our pipeline on the videos in the VoxCeleb2 dataset results in 8.8 million individual word utterances. This is massive and likely too large for practical use, resulting in 7.3 million positive case pairs and 18.9 trillion negative case pairs. This is however a preliminary test of the pipeline. There are limitations to vocabulary and visual aspects that will be applied in the future that will bring this number down to a more manageable level. It is important to note that there are many words that are only uttered once by an individual, thus the reason for fewer positive case pairs than total utterances. Also, the negative case count is only when comparing an individual's videos to themselves, not to other individuals' videos. Comparing individual to individual would result in the negative case count increasing astronomically.

This initial test does not filter out many cases that will be removed in published datasets. Some aspects that will be used in future works that publish dataset will be to remove cases that are of the same viseme (visually indistinguishable), remove word cases that only have one instance, ensure that the speakers head region of interest is large enough to perform authentication, ensure that the head position of the speaker is not to far rotated in any direction, limitations on word size, ensure balanced viseme coverage, etc. These methods will ensure that future datasets will be good representations of the world as well as remove many cases that would be unnecessarily difficult and result in very poor performance. These methods will also bring the size of the dataset down to make it more manageable compared to the size we found in the initial test of the pipeline.

Conclusion

Utilizing lip motion analysis as biometric authentication is a very promising field of research. Previous works in this area have proven that it has value and merit, but the datasets that have been used are generally so small that it is difficult to prove real-world generalization and applicability. We propose the solution to this issue is collecting much larger in-the-wild datasets using automated dataset extraction pipelines such as the one described.

In the wild datasets will enable and challenge future work in this area to improve upon what has been done and compare to other works. A benchmark dataset is required to compare the various methods in comparable works. The pipeline we propose will enable the collection of benchmark datasets and thus move the efforts forward to advance this new, promising form of biometric authentication.

References

[1] M. Zhou, Q. Wang, Q. Li, P. Jiang, J. Yang, C. Shen, C. Wang, and S. Ding, "Securing face liveness detection using unforgeable lip motion patterns," *arXiv preprint arXiv:2106.08013*, 2021.

[2] R. Shmelkin, L. Wolf, and T. Friedlander, "Generating master faces for dictionary attacks with a network-assisted latent space evolution," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 01–08.

[3] G. Chetty and M. Wagner, "Automated lip feature extraction for liveness verification in audio-video authentication," *Proc. Image and Vision Computing*, pp. 17–22, 2004.

[4] M. I. Faraj and J. Bigun, "Motion features from lip movement for person authentication," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 1059–1062.

[5] M.-I. Faraj and J. Bigun, "Audio–visual person authentication using lip-motion from orientation maps," *Pattern recognition letters*, vol. 28, no. 11, pp. 1368–1382, 2007.

[6] M. Ichino, Y. Yamazaki, W. Jian-Gang, and Y. W. Yun, "Text independent speaker gender recognition using lip movement," in *2012 12th International Conference on Control Automation Robotics & Vision (ICARCV)*, pp. 176–181.

[7] J. Chen, L. Cai, Y. Tu, R. Dong, D. An, and B. Zhang, "An identity authentication method based on multi-modal feature fusion," in *Journal of Physics: Conference Series*, vol. 1883, no. 1. IOP Publishing, 2021, p. 012060.

[8] A. B. Wong, "Authentication through sensing of tongue and lip motion via smartphone," in *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2021, pp. 1–2.

[9] F. Shafait, R. Kricke, I. Shdaifat, and R.-R. Grigat, "Real time lip motion analysis for a person authentication system using near infrared illumination," in *2006 International Conference on Image Processing*. IEEE, 2006, pp. 1957–1960.

[10] D. Shang, X. Zhang, and X. Xu, "Face and lip-reading authentication system based on android smart phones," in *2018 Chinese Automation Congress (CAC)*. IEEE, 2018, pp. 4178–4182.

[11] S. T. Kim and Y. M. Ro, "Attended relation feature representation of facial dynamics for facial authentication," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1768–1778, 2018.

[12] Z. Sun, D.-J. Lee, D. Zhang, and X. Li, "Concurrent two-factor identify verification using facial identify and facial actions," *Electronic Imaging*, vol. 2021, no. 6, pp. 318–1, 2021.

[13] Z. Sun, A. Sumsion, S. Torrie, and D.-J. Lee, "Learn dynamic facial motion representations using transformer encoder," in *2022 Intermountain Engineering, Technology and Computing (IETC)*. IEEE, 2022, pp. 1–5.

[14] Z. Sun, A. W. Sumsion, S. A. Torrie, and D.-J. Lee, "Learning facial motion representation with a lightweight encoder for identity verification," *Electronics*, vol. 11, no. 13, p. 1946, 2022.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[16] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre *et al.*, "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*, vol. 964. Citeseer, 1999, pp. 965–966.

[17] K. Ruengprateepsang, S. Wangsiripitak, and K. Pasupa, "Hybrid training of speaker and sentence models for one-shot lip password," in *International Conference on Neural*

Information Processing. Springer, 2020, pp. 363–374.
[18] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6219–6223.
[19] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz, "Robust face detection using the hausdorff distance," in *International conference on audio-and video-based biometric person authentication*. Springer, 2001, pp. 90–95.
[20] S. Lucey, "An evaluation of visual speech features for the tasks of speech and speaker recognition," in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 260–267.
[21] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.
[22] L. L. Mok, W. H. Lau, S. H. Leung, S.-L. Wang, and H. Yan, "Lip features selection with application to person authentication," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–397.
[23] A. Sayo, Y. Kajikawa, and M. Muneyasu, "Biometrics authentication method using lip motion in utterance," in *2011 8th International Conference on Information, Communications & Signal Processing*. IEEE, 2011, pp. 1–5.
[24] T. Nakata, M. Kashima, K. Sato, and M. Watanabe, "Lip-sync personal authentication system using movement feature of lip," in *2013 International Conference on Biometrics and Kansei Engineering*. IEEE, 2013, pp. 273–276.
[25] A. B. Hassanat, "Visual passwords using automatic lip reading," *arXiv preprint arXiv:1409.0924*, 2014.
[26] A. Basheer Hassanat, "Visual words for automatic lip-reading," *arXiv e-prints*, pp. arXiv–1409, 2014.
[27] Y. Yuan, J. Zhao, W. Xi, C. Qian, X. Zhang, and Z. Wang, "Salm: smartphone-based identity authentication using lip motion characteristics," in *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2017, pp. 1–8.
[28] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "Silentkey: A new authentication framework through ultrasonic-based lip reading," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–18, 2018.
[29] K. Lander and L. Chuang, "Why are moving faces easier to recognize?" *Visual Cognition*, vol. 12, no. 3, pp. 429–442, 2005.
[30] C. Sanderson and K. K. Paliwal, "Fast features for face authentication under illumination direction changes," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, 2003.
[31] M. U. R. Sanchez, *Aspects of facial biometrics for verification of personal identity*. University of Surrey (United Kingdom), 2000.
[32] C. Wright, D. Stewart, P. Miller, and F. Campbell-West, "Investigation into dct feature selection for visual lip-based biometric authentication," in *Irish Machine Vision & Image Processing Conference Proceedings*, 2015, pp. 11–18.
[33] C. Wright and D. Stewart, "One-shot-learning for visual lip-based biometric authentication," in *International Symposium on Visual Computing*. Springer, 2019, pp. 405–417.
[34] C. Wright and D. W. Stewart, "Understanding visual lip-based biometric authentication for mobile devices," *EURASIP Journal on Information Security*, vol. 2020, no. 1, pp. 1–16, 2020.
[35] S. A. Dar, S. Palanivel, M. K. Geetha, and M. Balasubramanian, "Mouth image based person authentication using dwlstm and gru," 2022.
[36] D. Estival, S. Cassidy, F. Cox, D. Burnham *et al.*, "Austalk: an audio-visual corpus of australian english," 2014.
[37] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103.
[38] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*. IEEE, 2019, pp. 1–8.
[39] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.
[40] J. S. Chung and A. Zisserman, "Lip reading in profile," 2017.
[41] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
[42] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.
[43] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett *et al.*, "Large-scale visual speech recognition," *arXiv preprint arXiv:1807.05162*, 2018.
[44] Y. Zhao, R. Xu, and M. Song, "A cascade sequence-to-sequence model for chinese mandarin lip reading," in *Proceedings of the ACM Multimedia Asia*, 2019, pp. 1–6.
[45] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
[46] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.
[47] J. Yuan, M. Liberman *et al.*, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.
[48] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

Author Biography

Shad Torrie received his B.E. degree in Computer Engineering with a minor in Computer Science from Brigham Young University in 2022. He is currently pursuing a Ph.D. degree in the Electrical and Computer Engineering Department at Brigham Young University. His research interests include Computer Vision, Deep Learning, Human Computer Interface and Robotics.

Andrew Sumsion received his B.E. degree in Electrical Engineering with minors in Computer Science and Math from Brigham Young University in 2022. He is currently a Ph.D. student in the Electrical and Computer Engineering Department at Brigham Young University. His research interests include Computer Vision, Deep Learning, and Robotic Vision.

Zheng Sun received his Bachelor of Engineering degree from Sun Yat-sen University in 2017. He is pursuing a Ph.D. degree in the Electrical and Computer Engineering Department at Brigham Young University. His work focuses on computer vision and machine learning.

Dr. D. J. Lee received his Ph.D. degree in electrical engineering from Texas Tech University in 1990 and MBA degree from Shenandoah University in 1999. He served in the machine vision industry for eleven years before joining Brigham Young University faculty in 2001. He is currently a professor and the director of the Robotic Vision Laboratory in the Electrical and Computer Engineering Department at BYU. He cofounded Smart Vision Works, Inc. in 2012. His research includes vision systems and devices with artificial intelligence, high-performance visual computing, real-time robotic vision, and visual inspection automation applications.