FACE IDENTIFICATION FROM LOW RESOLUTION NEAR-INFRARED IMAGES

Soumyadeep Ghosh, Rohit Keshari, Richa Singh and Mayank Vatsa

IIIT Delhi, India

ABSTRACT

Face identification from low quality and low resolution Near-Infrared (NIR) face images is a challenging problem. Since surveillance cameras typically acquire images at a large standoff distance, the effective resolution of the face is not large enough to identify the individuals. Moreover for a 24-hour surveillance footage, images in low light and at nighttime are acquired in NIR mode which makes the identification problem even more challenging. We propose an effective method using both hand-crafted and learned features for face identification of low resolution NIR images. We show that learned features contribute considerably to the performance of identification algorithm, and that using both feature level and score level fusion in a hierarchal approach gives good performance. The results demonstrate the effectiveness of the proposed approach on images which are of low quality, low resolution and acquired under challenging illumination conditions in near-infrared mode by surveillance cameras.

***Index Terms*—** Low resolution face recognition, Near Infrared, Autoencoder, Restricted Boltzmann Machine.

1. INTRODUCTION

Until the last decade, the primary usage of face recognition [1] was for access control which involved matching images of cooperative subjects in controlled environment. However, in the last decade, surveillance has become one of the foremost application areas for face recognition. A study [2] conducted by the British Security Industry Authority (BSIA) estimated that UK has has 5.9 million CCTV cameras, thus there is one camera for every 11 people. Generally surveillance cameras have large standoff distance which affects image quality and the effective resolution of the face is also low. Additional challenges such as illumination, expression and image quality makes the problem even more difficult. In literature this is known as face recognition at a distance (FRAD) [3]. Fig. 1 shows such sample surveillance quality images. The frame quality and recognition performance also depends on a lot of other factors such as the kind of camera used, depth of field, field of view of the camera, effective resolution and focal length of the camera. Surveillance cameras are fixed at a point and are used to capture images throughout day and night. Due to insufficient amount of visible light during nighttime, these cameras capture near infrared (NIR) spectrum images. Thus, the problem escalates to identification of individuals from low quality Near-Infrared face images acquired at very poor illumination conditions.

Several approaches have been proposed for matching visible to near-infrared face images. Most of them [5, 6, 7, 8, 9] used subspace and dictionary based techniques for matching visible (VIS) to NIR face images. Yi et al. [10] trained a multimodal learning based model to learn shared representation of VIS and NIR images and utilized it for cross spectral face recognition.

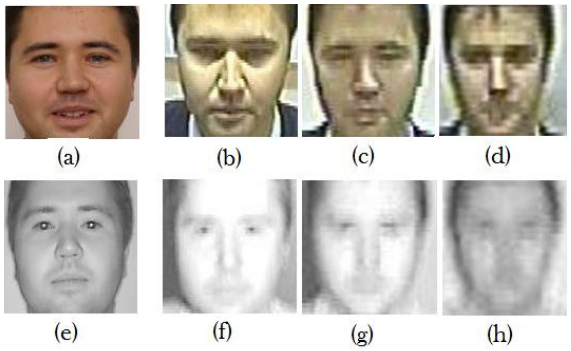


Fig. 1: Images in the SCface database [4], (a) Visible gallery image, (b)-(d) visible spectrum images acquired from surveillance cameras from different standoff distances, (e) NIR mugshot image, (f)-(h) NIR images acquired from surveillance cameras from different standoff distances

.

On the other hand, there have been several methods in literature for cross-resolution face recognition. Bhatt. et al. [11] proposed a co-transfer based learning approach for matching faces across resolution. Biswas et al. used multidimensional scaling [12] for face recognition across pose, illumination, and resolution. In these works, results were shown on surveillance quality images in visible spectrum. However, there are a very few approaches which have attempted to provide a solution to both these problems in a single algorithm. The only method known to the best of our knowledge is by Kang et al. [13], which provided a restoration based approach for matching visible and NIR face images taken at nighttime. An extensive camera setup was used to acquire the images at different distances during both daytime and nighttime. They used Locally linear embedding [14] to restore the long distance images, and matched using an existing heterogeneous face matching algorithm [15]. Most of the face recognition algorithms for low quality images have used sophisticated restoration and preprocessing methods prior to recognition [16]. The near-infrared face recognition at a distance (NFRAD) [17] problem deals with both cross spectral and cross distance face matching. To the best of our knowledge there is no work in literature which has explored the effectiveness of unsupervised feature learning for this problem.

1.1. Research Contribution

The research contribution of this paper is two-folds.

- Propose an algorithm which combines learned and hand-crafted features for cross-resolution near infrared face recognition.
- Without performing any preprocessing or enhancement for

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

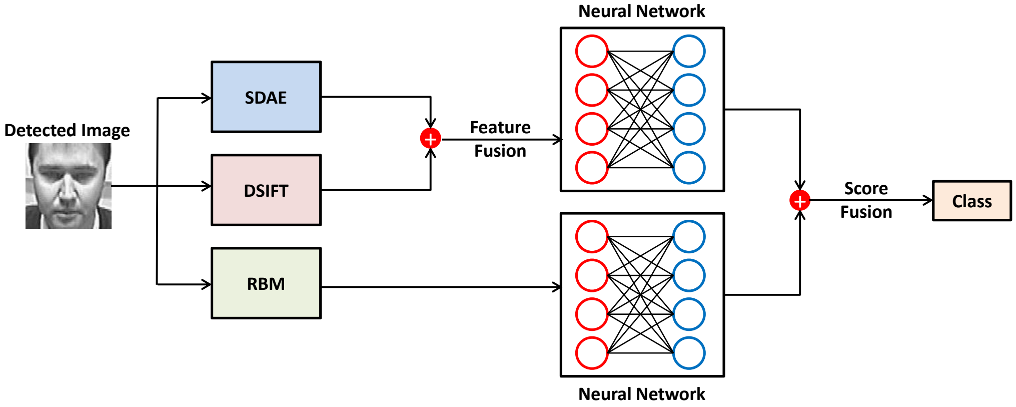


Fig. 2: Steps involved in the proposed method showing hierarchical fusion at both feature and score level.

the low quality images, the algorithm yields state-of-the-art results on the SCface database (cross spectral cases).

2. PROPOSED ALGORITHM

In this research, we propose a face recognition algorithm which is a combination of learnt and handcrafted features. As shown in Fig. 2, the features are learned using Stacked Denoising Autoencoder (SDAE) [18] and Restricted Boltzmann Machine (RBM) [19]. Along with these two, DSIFT [20] features are extracted that provide histogram of oriented gradients at densely sampled keypoints. It is also shown that performing feature level fusion of learned features with dense SIFT yields superior identification accuracies. Finally, an efficient architecture is proposed which uses both feature level and score level fusion at different levels.

2.1. Feature Extraction

In literature, histogram of Oriented gradient based features (D-SIFT) have been shown [8] to be effective in visible (VIS) to NIR face recognition. SIFT [20] features were also shown [11, 12] to be effective in low resolution face recognition. Therefore we have extracted dense SIFT features from 72 × 72 images. Each keypoint gives a descriptor of size 128, all the descriptors from the keypoints of an image are concatenated to produce the descriptor of the entire image. Along with D-SIFT, two learnt representations are also extracted.

2.1.1. Stacked Denoising Autoencoder

A classical autoencoder learns a function f_θ which is a mapping of the input data $x'$to a representation y given by

$$y = f_\theta(x') = s(Wx' + b) \tag{1}$$

known as the encoder, where x' is the corrupted version of x. The decoding part deals with the reconstruction of x' given by

$$\hat{x} = g_{\theta'}(y) = s(W'y + b') \tag{2}$$

where $\hat{x}$ is the reconstructed data. The objective function is to minimize the reconstruction error given by

$$L_\theta(x, \hat{x}) = argmin_\theta ||x - \hat{x}||_2 + \frac{\lambda}{2}||W||_2 \tag{3}$$

where $\frac{\lambda}{2}||W||_2$ is the L_2 regularization term which prevents overfitting by performing weight decay.

2.1.2. Restricted Boltzman Machine

The Restricted Boltzman Machine (RBM) [19] is an undirected graphical model which has two layers, namely visible layer $V \in \{0,1\}^A$ and the hidden layer $h \in \{0,1\}^B$, containing stochastic units. In the model, each unit in the visible layer is connected to each unit in the hidden layer. RBM models the following energy function $E : \{0,1\}^{A+B} \to R$ defined as

$$E(v, h; \theta) = -\sum_{i=1}^{A}\sum_{j=1}^{B} v_i w_{i,j} h_j - \sum_{i=1}^{A} b_i v_i - \sum_{j=1}^{B} a_j h_j \tag{4}$$

where the model parameters are $\theta = a, b, W$. The joint distribution of the hidden and visible layers is given by:

$$P(v, h; \theta) = \frac{1}{z(\theta)} exp(-E(v, h; \theta)) \tag{5}$$

The above formulation is for a binary RBM, when the visible units are real values we use a Gaussian RBM [19]. This model is pre-trained on a large set of data and then fine-tuned on target data which makes feature extraction by this model efficient.

2.2. Classification and Fusion

As illustrated in the block diagram (Fig. 2), DSIFT and SDAE features are normalized and then concatenated to create a combined feature vector. Two 2-layer neural networks are learnt, one for the concatenated feature vector and another for RBM. The neural networks are trained using stochastic gradient descent algorithm to generate the match scores. The match scores are further normalized and then combined using weighted sum rule for final classification/decision.

3. EXPERIMENTS AND ANALYSIS

Since surveillance primarily requires identification with low resolution videos captured in visible and NIR spectrums, the efficacy of the proposed algorithm is evaluated in identification mode. The performance of the proposed framework is evaluated using the SCface database [4] which contains low resolution face images in both NIR and visible spectrum.

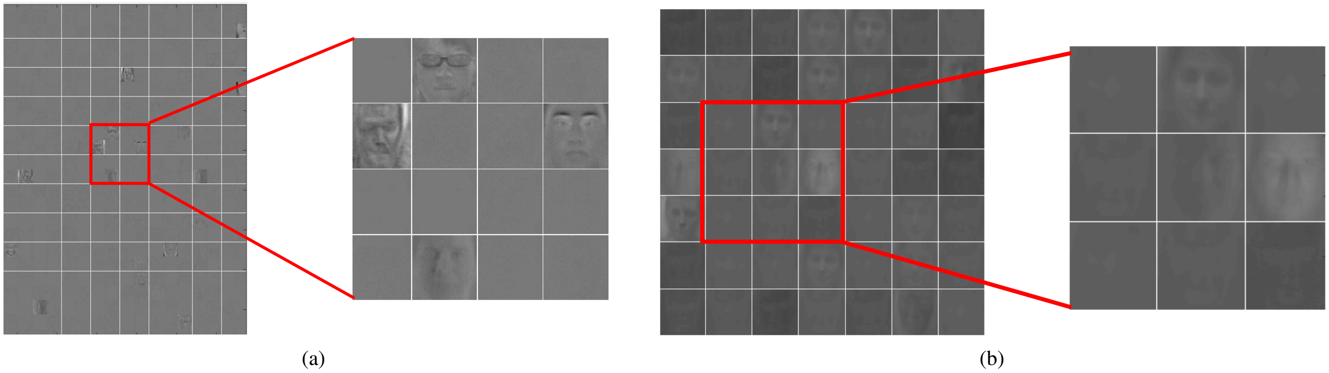


Fig. 3: Visualization of the learned weights of (a) RBM and (b) layer 1 of SDAE after 100 epochs of training

3.1. Dataset and Protocol

The SCface database [4] has 4160 images of 130 subjects acquired using 8 different surveillance cameras mounted at different angles and 3 different standoff distances: 1, 2.6 and 4.2 meters. Each camera produces images of 3 different resolutions depending on the standoff distance. The closest standoff distance produces image of resolution 72 × 72, followed by 48 × 48 and 32 × 32. Out of the 8 cameras, 3 cameras operate in NIR mode, using which images are acquired at challenging illuminating conditions at nighttime. We use 72 × 72 and 48 × 48 images of 6 cameras (5 visible spectrum cameras and 1 NIR camera) for training the neural network classifier. Since stacked denoising autoencoder and restricted Boltzman Machine require large amount of data for learning representations, they are pre-trained on 50, 250 face images of the CMU MultiPIE database [21], and fine-tuned on the 72 × 72 and 48 × 48 images of these 6 cameras of the SCface database. Images from the other 2 NIR cameras with 48 × 48 and 32 × 32 resolution are used for testing in identification mode. As can be seen in Fig. 1, such images are of extremely low quality and the faces in 32 × 32 NIR images are barely recognizable.

3.2. Experiments

Faces are detected by using the Viola Jones face detector [22]. Since the quality of low resolution NIR videos is very poor, Adaboost based face detector is not able to extract all the faces from the frames. Therefore, in such cases, faces are manually detected. For extracting DSIFT features, we use 25 keypoints in each image, aligned on an uniform grid, keeping the spatial bin size to 12 pixels. Thus the dimensionality of DSIFT features for each image is 3200. For training the SDAE, we use two layer network having 4096 and 2048 nodes in the first and second hidden layers respectively. For L_2 regularization used in the SDAE the value of λ is 0.3 . The trained RBM has 2048 nodes in the hidden layer. The neural network classifier has 2 hidden layers having 512 and 256 nodes in them. The neural network has 130 nodes in the output layer which is equal to the number of subjects in the database.

We have compared the performance of the proposed architecture (Fig. 2) with several other architectures namely, using raw pixels, DSIFT, SDAE and RBM as features, feature fusion by combining these features and COTS (FaceVacs [23]). For each of these architectures we have trained a neural network classifier and the classification results are summarized in Table 1.

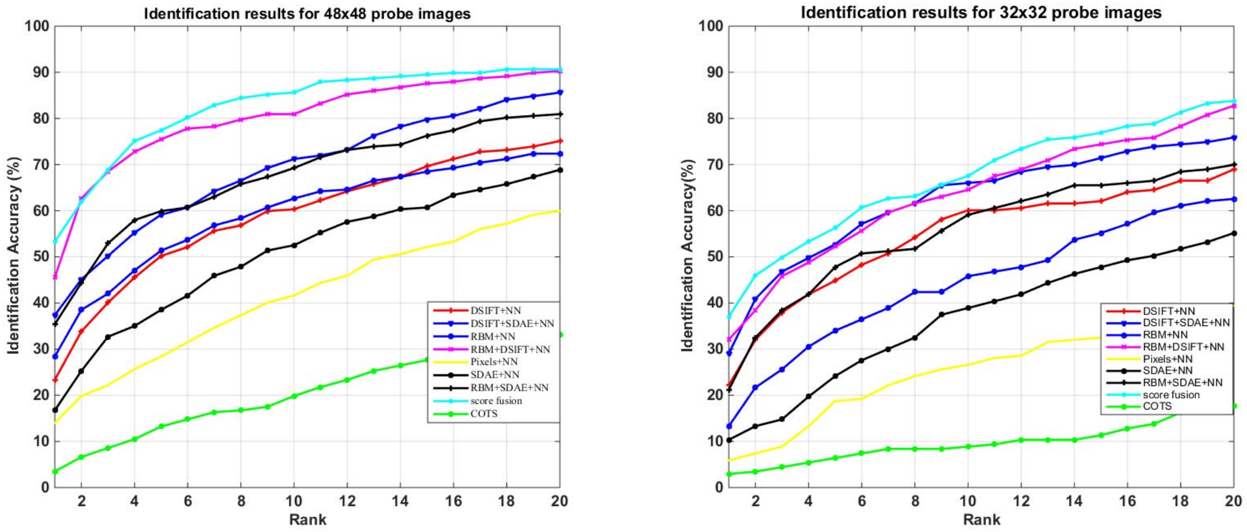
Table 1: Rank 1 identification accuracies of different algorithms on different resolutions of the NIR probe images.

Algorithm	Rank 1 accuracy (%)	
	48 × 48	32 × 32
FaceVacs [23] (COTS)	3.50	2.94
Pixels + Neural Network(NN)	14.03	5.91
DSIFT + NN (A)	23.34	22.16
SDAE + NN (B)	16.73	10.34
RBM + NN (C)	28.40	13.30
RBM + SDAE + NN (D)	35.41	21.18
(SDAE + DSIFT) + NN (X)	37.35	29.06
(RBM + DSIFT) + NN (Y)	45.52	32.07
Score Fusion of (C) and (X)	**53.81**	**37.03**

3.3. Analysis of Results

The identification results obtained after performing the above experiments are given in Table 1 and CMC curves are shown in Fig. 4, for both 48 × 48 and 32 × 32 probe images. The analysis of the experimental results are as follows:

- As shown in previous research [8], DSIFT yields competitive accuracies on this problem. RBM features perform better than DISFT features for 48 × 48 probe images, which shows that unsupervised learned features are more effective than handcrafted features which has been extensively used by recent methods [12, 11] for face recognition on surveillance quality images. It can also be seen that RBM features perform better than SDAE features for both 32 × 32 and 48 × 48 images. It can be observed from the visualization of learned weights of RBM and SDAE (Fig. 3) that the representation learned by the former are much better than the latter. The experimental results (Table 1) also correlate with this observation.
- The feature level fusion of RBM/SDAE features with D-SIFT yields increment in identification accuracies. This shows that fusion of handcrafted and learned features are effective. Fusion at the score level gives further improvement in identification accuracies. Using both feature level and score level fusion gives the best results for both 32 × 32 and 48 × 48 images.
- FaceVacs [23], a commercial-off-the-shelf (COTS) system



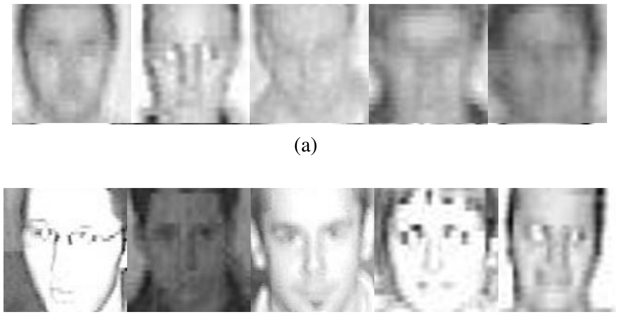
(a)

(b)

Fig. 4: Identification accuracies of existing algorithms and different variants of the proposed algorithm on the SCface database when probe images are of size (a) 48×48 and (b) 32×32.

yields lower accuracies compared to the proposed algorithm. The COTS yields rank-1 identification accuracy of 2.94% and 3.50% for 32×32 and 48×48 NIR probe images respectively. This demonstrates that even high performing commercial face recognition systems cannot recognize low resolution images efficiently.

- Fig. 5 shows some of the images that were incorrectly classified and correctly classified. It can be seen that the images that were not classified correctly are generally of extremely low quality. However, some of the poor quality images were correctly classified as well as shown in Fig. 5(b). It can be observed that if the image quality is poor with minor presence of other covariates such as pose, then the image is generally correctly classified. However, very poor quality images as shown in Fig. 5(a) are difficult to recognize. It is worth mentioning unless there is familiarity, these images are difficult to be matched by humans as well.



(a)

(b)

Fig. 5: Sample probe images demonstrating the results. (a) incorrectly classified probe images, (b) correctly classified probe images.

3.4. Running time

The unsupervised feature extractors are pretrained on MultiPIE [21] database which is performed in 2.1 hours for the RBM and 4.2 hours for the SDAE to train. After feature extraction, training the neural network classifier is performed in considerably less time. For example, training two layer neural network classifier for SDAE+DSIFT features is performed in 4.2 minutes. During testing, the speed of classification is about 1613 images per second. All implementations are perfromed on a desktop with 3.4 GHz Intel Core i7 processor and 16GB of RAM.

4. CONCLUSION

We propose an effective method for identification of low resolution, low quality NIR images acquired by surveillance cameras under challenging illumination conditions from varying standoff distances. Of all the different architectures studied, the best performing one uses both feature level fusion and score level fusion. It utilizes the effectiveness of using both hand-crafted and learned features. Identification results on the SCface database demonstrate the effectiveness of the algorithm compared to a commercial algorithm and existing algorithms. In future, we plan to explore other algorithms for learning features to further improve the performance along with analyzing the effect of preprocessing on recognition performance.

5. ACKNOWLEDGEMENT

This research is supported in part by a grant from the Department of Electronics and Information Technology, Government of India.

6. REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] "One surveillance camera for every 11 people in britain," http://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html, Accessed: 2016-01-29.

[3] A.K Jain and S.Z Li, *Handbook of face recognition*, vol. 1, Springer, 2005.

[4] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Springer MTA*, vol. 51, no. 3, pp. 863–879, 2011.

[5] D. Yi, R. Liu, R. Chu, Z. Lei, and S.Z Li, "Face matching between near infrared and visible light images," in *IAPR ICB*, pp. 523–530. 2007.

[6] J-Y Zhu, W-S. Zheng, J-H. Lai, and S. Z Li, "Matching nir face to vis face using transduction," *IEEE TIFS*, vol. 9, no. 3, pp. 501–514, 2014.

[7] S. Ghosh, T. I Dhamecha, R. Keshari, R. Singh, and M. Vatsa, "Feature and keypoint selection for visible to near-infrared face matching," in *IEEE BTAS*, 2015, pp. 1–7.

[8] T.I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, "On effectiveness of histogram of oriented gradient features for visible to near infrared face matching," in *IAPR ICPR*, 2014, pp. 1788–1793.

[9] F. Juefei-Xu, D. Pal, and M. Savvides, "Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *IEEE CVPR*, 2015, pp. 141–150.

[10] D. Yi, Z. Lei, and S.Z Li, "Shared representation learning for heterogenous face recognition," in *IEEE FG*, 2015, vol. 1, pp. 1–7.

[11] H.S. Bhatt, R. Singh, M. Vatsa, and N.K. Ratha, "Improving cross-resolution face matching using ensemble-based co-transfer learning," *IEEE TIP*, vol. 23, no. 12, pp. 5654–5669, 2014.

[12] S.P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE TPAMI*, vol. 38, no. 5, pp. 1034–1040, 2015.

[13] D. Kang, H. Han, A.K. Jain, and S-W. Lee, "Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching," *PR*, vol. 47, no. 12, pp. 3750–3766, 2014.

[14] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[15] B.F Klare and A.K Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE TPAMI*, vol. 35, no. 6, pp. 1410–1422, 2013.

[16] S. Bharadwaj, M. Vatsa, and R. Singh, "Biometric quality: a review of fingerprint, iris, and face," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–28, 2014.

[17] H. Maeng, H-C. Choi, U. Park, S-W. Lee, and A.K. Jain, "Nfrad: Near-infrared face recognition at a distance," in *IEEE IJCB*, 2011, pp. 1–7.

[18] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *JMLR*, vol. 11, pp. 3371–3408, 2010.

[19] R Salakhutdinov, A Mnih, and G Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 791–798.

[20] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[21] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "The CMU multi-pose, illumination, and expression (Multi-PIE) face database," Tech. Rep., Carnegie Mellon University Robotics Institute. TR-07-08, 2007.

[22] P. Viola and M.J. Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.

[23] "FaceVacs Technology," http://www.cognitec.com/technology.html, Accessed: 2016-02-07.