Review of Adult Image Detection Methods

Sasan Karamizadeh
Iran Telecommunication Research Center, Tehran, Iran, s.karamizadeh@itrc.ac.ir

Abouzar Arabsorkhi
Iran Telecommunication Research Center, Tehran, Iran

Abstract

Nowadays, adult images and other such indecent matter are available on the social media and the Internet for children. Filtering of adult image has become one of the big changes for searches; they are tied to finding methods to filter adult images. Social media network is interested in filter adult images from normal ones. Analysis method uses the bright image to automatically detect and filter images in the media. In this paper, we have reviewed methods such as color based, shape based, local and global feature approach, deep learning and bag-of-words for filtering adult images which include comparing with the advantages and disadvantages.

Keywords: *Internet, filtering, adult image, deep learning, shape based*

1. Introduction

Filtering complex media such as violent, pornographic has become significant because of the intense utilization of online media by individuals of any age; and among delicate media sorts, pornography entertainment is *regularly* the most unwelcome. A kind of utilizations has expanded societal enthusiasm on the issue, e.g., recognizing unseemly conduct by means of observation cameras; or diminishing the trading of sexually-charged texts, otherwise called "sexting", by minors [1]. Furthermore, law implementers may utilize explicit entertainment channels as a first filter when searching for teenager pornography entertainment in the criminological examination of computers or internet content. There are two major methods of identifying images related to porn namely i) concentrating on web page contents with that picture that will arrange the picture content alternately; ii) they search inside the images [2]. They judge these contents using skin texture pixels [3]. In this paper, we reviewed several pornography filtering detection methods.

2. Procedure for Paper Submission

The color-based approach depends on the activity suspicion that pixels in pornographic images are mostly skin [1], [2]. Color models are normally used to characterize a vigorous skin color representation [3].Skin proportion, histogram, color probabilities, and associated segments are utilized to describe skin color. The most vital feature used to control an image that contains porn or otherwise is color [4]. Skin filter is one of the important parts of a porn detection system, which *expects* to determine which pixels are skin colored or not utilizing a skin color appropriation model [5], [6]. To classify the skin pixels as or otherwise not of skin color, the models on the distribution of skin utilize a space with a single color [7], [8]. This is built on the result to show the selection of space's color is insignificant if a thresholding is carried using enough training as well as a proper false alarm. It can be divided into three subcategories: color range computed color histograms, and parametric color distribution functions [9], [10].

2.1 Computed Color Histograms

There are two *problems* that must be tackled in order to build a histogram model: the choice of the size of the histogram and color space, which is measured by the number of channels in the color channel [11]. Color histogram is verified for small databases. Only color information is recorded by a color histogram; images can be completely different with similar color histograms [12]. Minor computing and robustly tolerant movement of an object in an image are advantages of the color histograms [11]. A color

histogram is a vector where every passage stores the quantity of pixels of a given color in the image [11]. Histogramming is applied after all the images have been scaled to contain a similar number of pixels. Colors are mapped into a discrete color spade space containing colors [12], [13].

3. Shape Based

Shape is one of the important characteristics of an object. Shapes have been extracted from skin regions in the shape-based method. The exclusively describe the object shape is gold of shape descriptor. Shape descriptor has to minimize the maximize the between-class variance and be insensitive to noise [14]. Boundary, region and moment have represented in shape [14], [15]. For making measurements of shapes, or recognition objects and matching shapes can be utilized these representations. The contour-based feature, Geometric constraints, Gaussian model and Hu and Zernike moments of the skin distribution are four categories of shape based [16], [17].

4. Contour-Based Feature

Boundary information is extracted by using contour based features techniques. The contour based shape can be separated into structure and global methods [18]. The global method does not divide the shape into subparts to drive the feature vector; a matching process is used to complete the boundary information, therefore known as the continuous approach [19]. The shape is broken down using a structural approach into subparts, and this approach is known as the separate method. Normally, the structural approach is represented by a string or tree and graph, which is finally used to reconcile the image retrieval process [20].

5. Shape Descriptors

To exclusively characterize the object shape is one of the aim of shape descriptors [21]. Shape descriptor have to maximize the between class variance, minimize the within-class variance and insensitive to noise. Three types of Shape descriptor are utilized which are i) three simple which is divided to rectangularity, *compactness* and irregularity. irregularity is distance ratio between the major and minor axes of the objects. ii) Seven normal moment invariants that is independent of scale, translation and rotation. Iii)Zernike moments which is set of Zernike polynomials defined over the polar coordinate space inside a unit circle [22].

6. Local and Global Feature Approach

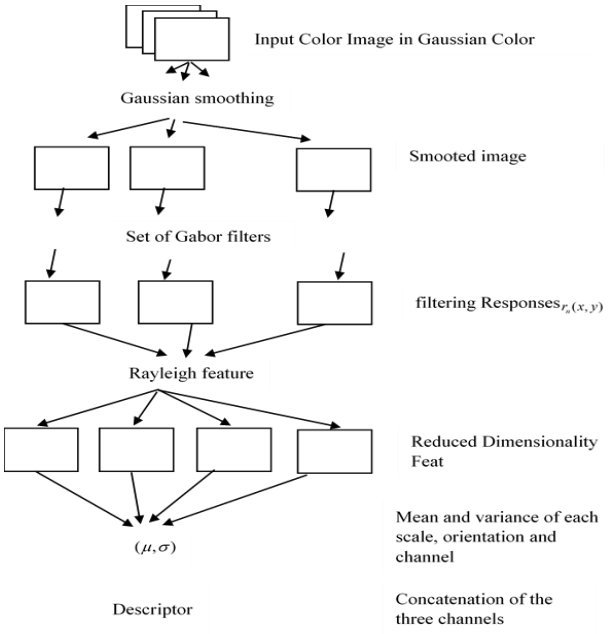

Figure 1. Color texture description scheme.

Essentially, the local and global features contain two types of feature extractions from the image based on application. Usually, image retrieval is utilized in global descriptors, classification, and *object* detection, whereas local is used for the object [23]. For classifying normal and pornographic images, a local description using a local feature approach is applied [24].

Whole image is described by global features to generalize the complete object while image patches are used in local features for the description of an object. Texture in an image patch is represented by the local features. Contour representation, texture feature, and shape descriptors are part of the global features. The feature method uses the Scale Invariant Feature Transform (SIFT) as well as the *Probabilistic* Latent Semantic Analysis (PLSA) [14], [15], [25]. Figure 1 shows global and local image descriptors

7. Deep Learning

Computational model is used by deep learning to compose multiprocessing layers to learn representations of data with multiple levels of abstraction [26]. Most neural networks exist due to the availability of data and the cost of computation is a single layer. To solve pattern recognition problems, the deep learning technique is utilized, which uses multiple layers for learning in the neural networks. Deep learning can be divided into four categories. Multiple levels of representation are utilized by the deep-learning method to represent the learning method [27].

7.1 CNN-Based Methods

One of the famous deep learning methods is the Convolutional Neural Networks where multiple layers are trained in a strong manner. This method is utilized by under the computer vision because it has been greatly effective [16]. The Convolutional layer is used by several kernels to convolve the complete image in addition to the intermediate creating numerous feature maps; pooling *layer* follows a convolutional layer for reducing overfitting and a fully-connected layer flow, finally the network layers for classification [17], [18].

7.2 AlexNet Architecture

This architecture is introduced by Alex Krizhevsky that was utilized to win the 2012 ILSVRC. This architecture has *been* made up of 5 convolutional layers, max-pooling layers and dropout layers, and finally 3 fully connected layers which is utilized for classification with 1000 possible categories. Figure 2. Shows the AlexNet Architecture [19].

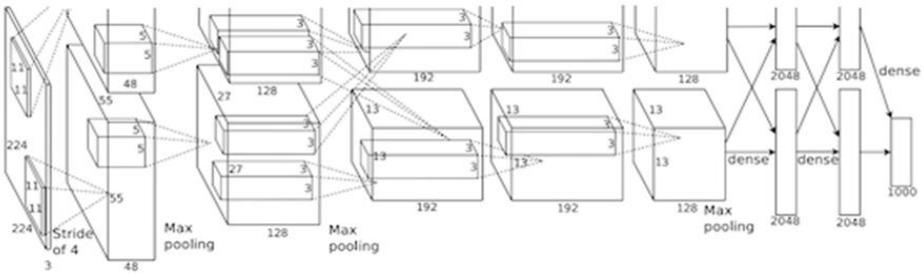

Figure 2. Shows Alexnet architecture.

7.3 GoogleNet Architecture

This architecture was introduced by google which is 22 layer CNN. One of important method to improve performance on deep learning is utilize more data and more layers. GoogleNet has been utilized 9 inception modules. One of disadvantage is that use more parameters models income more overfit [20]. Figure 3. shows googleNet architecture.

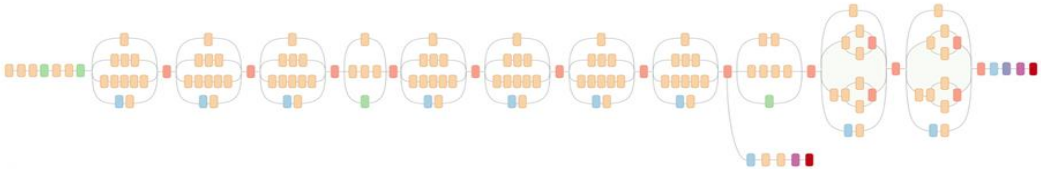

Figure 3. Shows Googlenet architecture.

7.4 VGG Architecture

This Architecture was introduced by Simonyan and Zisserman which is using 3 convolutional layers and max pooling *for* reducing size and finally 2 fully- connection layers are followed by a softmax classifier [28]. Figure 4 shows VGG architecture.

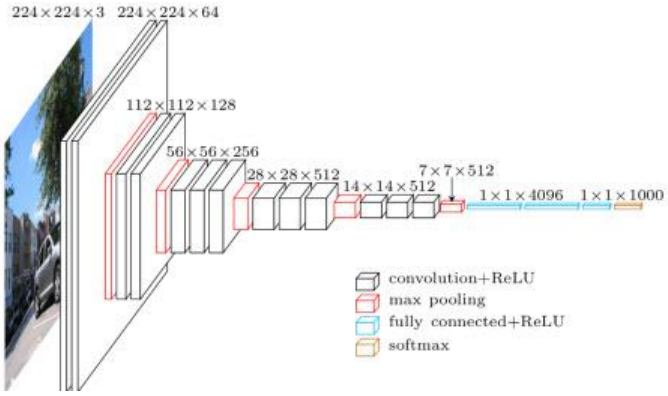

Figure 4. Shows VGG architecture.

7.5 Auto Encoding-Based Methods

Different kinds of artificial neural networks are utilized to study effective encoding, which includes the auto-encoder [29]. Substitute for preparing the system to foresee some objective esteem Y given sources *of* info X, an auto-encoder is prepared to remake the inputs itself. Thus, the output vectors contain similar dimensions as the input vector. Normally, a single layer has been unable to receive representative feature and act discriminatively from the raw data [23]. An Auto encoding Deep auto-encoder is utilized; this then pushes forward the code, which sends the learned code from the past auto-encoder to the following one to achieve its tasks. Sparse auto-encoding is one of supervised methods to learn feature from the unlabeled data automatically [24], de noising auto encoding makes the recovery with correct input from the ruined sort [30]; and contractive auto encoder: robustness is achieved through the addition of the penalty on analytic contractions to the reconstruction error's function. [31] are the sub-categories of auto encoding.

7.6 Sparse Coding-Based Methods

One of the unsupervised methods for learning sets of over complete basis to represent data comfortability is sparse coding. Basic vectors are able to capture inherent patterns and structures in the input data by having an over-complete basic. Sparse coding includes i) sparse coding SPM, which has been widely applied in the processing of sensory data, such as acoustic and image signals [32]; ii) *local* coordinate coding - The main idea is to locally embed points on the manifold into a lower dimensional space, expressed as coordinates with respect to a set of anchor points [26]; iii) super-vector coding which is algorithmically a simple extension of Vector Quantization (VQ) coding [27]; and iv) laplacian sparse coding where the same features are not just provided for the assigned to cluster centers that are optimally-selected, which again guarantees the chosen cluster centers to be the same [33], [34].

7.7 RBM-Based Methods

Single layer of hidden units is an RBM with no connection to each other with undirected, regular connections to a layer of visible units. Created date is stated from the RBM using a starting random state in one of the layers and then performing an alternating Gibbs sampling [35]. All units in one have been updated with parallel given to the present states of the units in the other layer. This will be repeated until the system samples from its distribution equilibrium. This can be divided into category deep belief networks, which includes a model that is probabilistic and generative that offer a joint distribution of probability over the absorbable labels as well as data [36], deep Boltzmann machines include another learning algorithm where the units are arranged in layers [37], and the mode, swotj with energy models are deep with just a single layer of stochastic hidden units to perform efficient training as well as inferences [38].

Table 1 shows advantages and disadvantages of four categories in deep learning. Normally, Generalization is used to show effective approach in diverse media such as images, texts, and audio; 'Unsupervised *learning'* mention about ability for learning a deep model without supervisory a notation. 'Feature learning' is mention learn feature based on a dataset without training. 'Biological understanding' shows whether this approach is basically a biological basis or theoretical basis. 'Small training set' mention to capability of a small number of example to learn a deep model.

Table 1. Comparison among Categories of Deep Learning

PROPERTIES	RBMS	SPARSE CODING	CNNS	AUTO-ENCODER
Generalization	✓	✓	✓	✓
Unsupervised learning	✓	✓	×	✓
Feature learning	✓	×	✓	✓
Biological understanding	×	✓	×	×
Small training set	✓	✓	✓	✓

8. Bag-of-Words(BoW) Method

One of the supervised models is the BoW. The human has great skills to reduce the computational cost of a data-driven visual search by attention mechanism. Bottom-up and top-down are two ways *to* direct attention [39]. In a bottoms-up method, for deployment of visual attention detecting salient regions in an image is utilized. Past knowledge is made available using prior knowledge availability for a specific target to guide the visual attention utilized by top-down way.

The advantage of Bow can be divided into i) mainly not affected by the position of the object being destructed in an image; position the object's destruction in an image, ii) Fixed length vector regardless of the amount of detections, iii) Successful in tabulating images based on the object they contain, vi) Still needs more examination for large changes in the scale and viewpoint. Shape of *visual* word position is not obvious and Localization objects are poor in images and a disadvantage to BoW. The bag-of-words method has been recommended in domain challenge text recovery for proposed in the text recovery domain problem for text document analysis [40]. BoW model is utilized by visual analogue of a word that is developed in the process of the vector quantization by clustering low-level visual aspects of points or the localized regions including the texture as well as the color [41]. BoW includes some steps that extract the features as mentioned in Table 2 below:

Table 2. BoW Model Steps

I.	**Automatic detection of the points/regions of interest;**
II.	Compute the descriptors that are local across the points/region;
III.	Quantify the descriptors into words that outline the visual vocabulary; and
IV.	Look for the image's occurrences in each particular word in the vocabulary to construct the feature of the Bag of Words.

9. Discussion

The color-based approach depends on the light of the activity suspicion that pixels in pornographic images are primarily skin. Color is the most straightforward visual quality to model, and it is a characteristic beginning stage when working with huge data sets. Three-dimensional color spaces bring about computationally reasonable calculations and algorithms that can be simply visualized. Global descriptor is used by shape. Texture, local pattern, and color have not been detecting the internal details with Shape. Thus, the classification of objects is used by shape. It cannot be utilized to recognize two objects of the same shape with various texture or colors. Therefore, you cannot utilize shape feature for object recognition.

Shape-based recovery will likewise basically represent the query and database images in some feature space, and endeavor Map and Reduce operations for discovering separations from the query image and restore the cases with minimum distance.

For improving the detection performance, one typically uses shape and skin color together. Global feature is *utilized* for low-level application such as classification and object detection. Local feature is applied for higher level application for example object recognition. To improve the accuracy of recognition, the Integration for local and global features was used.

Neural network is interpreted by these architectures. Level of network hardwiring is the Key difference between auto-encoder and CNN. Convolutional nets are hardwired. Convolution operation much have a connection view by sparsity in the number of connection in the neural network. Pooling (subsampling) operation in image space is likewise a hardwired set of neural connection in the *neural* domain. Usually, the CNN is used for image and speech task utilized.

Particularly the topology of the network is not specified by auto-enders. To rebuild the input, it is good knowledge to find great neural transformation. The decoder and encoder compos reconstruct the input. Latent factors or latent features are learned by a hidden layer.

Interpretation of the neural network is different with RMB. A two-part graph is interpreted by RMBs where knowledge is the learning from the distribution of hidden and input variables. CNN and Auto encoder have learned a function. Additionally, the model is generative by RBMs. The sample is created from a learned hidden representation. RBMS are trained by different algorithms. However, *at* the end of the day, after learning RBMs, you can use its network weights to interpret it as a feedforward network.

Sparse coding is *taught* to a general class of methods that automatically select a sparse arrangement of vectors from the expansive pool of conceivable bases to encode an input signal. Sparse coding can simply be connected to the BoW framework as a substitution for vector quantization. This approach claims that the utilization of sparse coding to develop abnormal state highlights, demonstrating that the resulting representation performance is better than the conventional representation. Utilization the sparse coding to develop abnormal state highlights demonstrates that the resulting representation performance is better than the conventional representation.

10. Conclusion

This paper has been presented a review of methods for filtering adult image. It divides them into six categories: color base which is color is the most straightforward visual quality to model, and it is a *characteristic* beginning stage when working with huge data sets. Shape based recovery will likewise basically represent the query and database images in some feature space, local and global, feature approach is utilized for low-level application such as classification and object detection. Bag of words is utilized to image classification, treating image features as words and deep which is utilized Computational model by deep learning to compose multiprocessing layers to learn representations of data with multiple levels of abstraction. The paper mainly reports the developments of the filtering methods. Each of them is included in some categories. The advantage and disadvantage are discussed in this paper.

Acknowledgment

This research has been supported by Iran Telecommunication Research Center and this review was as part of the applied research activities of the Socio-cultural Protection Plan using intelligent systems in the Information and Communications Technology Research Institute

References

[1] M. Alizadeh, W. H. Hassan, N. Behboodian, and S. Karamizadeh, "A brief review of mobile cloud computing opportunities," *Research Notes in Information Science* , vol. 12, pp. 155-160, 2013.

[2] R. Alvear-Sandoval and A. R. Figueiras-Vidal, "Does diversity improve deep learning?" presented at 2015 IEEE 23rd European Signal Processing Conference (EUSIPCO), Nice, France, August 31-September 4, 2015.

[3] A. Aurisano, A. Radovic, D. Rocco, A. Himmel, M. Messier, E. Niner, G. Pawloski, F. Psihas, A. Sousa, and P. Vahle, "A convolutional neural network neutrino event classifier," *Journal of Instrumentation*, vol. 11, no. 9, p. 09001, 2016.

[4] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. ICML Workshop on Unsupervised and Transfer Learning*, 2012.

[5] S. S. Chaeikar, A. B. A. Manaf, and M. Zamani, "Comparative analysis of master-key and interpretative key management (IKM) frameworks," in *Cryptography and Security in Computing*, Croatia: InTech, 2012.
[6] S. S. Chaeikar, M. Zamani, C. S. Chukwuekezie, and M. Alizadeh, "Electronic voting systems for European Union countries," *Journal of Next Generation Information Technology*, vol. 4, no. 5, p. 16, 2013.
[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *Computer Science*, 2014.
[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *Workshop on Statistical Learning in Computer Vision ECCV*, vol. 44, no. 247, pp. 1-22, 2004.
[9] K. Dong, L. Guo, and Q. Fu, "An adult image detection algorithm based on bag-of-visual-words and text information," in *Proc. 2014 10th International Conference on Natural Computation (ICNC)*, IEEE, 2014.
[10] L. Duan, G. Cui, W. Gao, and H. Zhang, "Adult image detection method base-on skin color model and support vector machine," in *Proc. Asian Conference on Computer Vision*, 2002.
[11] S. Gao, I. W. H. Tsang, L. T. Chia, and P. Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *Proc. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2010.
[12] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. (2017). A review on deep learning techniques applied to semantic segmentation. [Online]. Available: https://arxiv.org/pdf/1704.06857.pdf
[13] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27-48, 2016.
[14] W. Hsu, S. Chua, and H. Pung, "An integrated color-spatial approach to content-based image retrieval," in *Proc. the Third ACM International Conference on Multimedia*, ACM, 1995.
[15] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," *Journal of Signal and Information Processing*, vol. 4, no. 3, p. 173, 2013.
[16] C. Y. Jeong, J. S. Kim, and K. S. Hong, "Appearance-based nude image detection," in *Proc. the 17th International Conference on Pattern Recognition, ICPR 2004*, IEEE, 2004.
[17] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81-96, 2002.
[18] F. Karamizadeh, "Face recognition by implying illumination techniques–A review paper," *Journal of Science and Engineering*, vol. 6, no. 1, pp. 1-7, 2015.
[19] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," *Journal of Signal and Information Processing*, vol. 4, no. 3, p. 173, 2013.
[20] S. Karamizadeh, S. M. Abdullah, J. Shayan, P. Nooralishahi, and B. Bagherian, "Threshold based skin color classification," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 2-3, pp. 131-134, 2017.
[21] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527-1554, 2006.
[22] G. E. Hinton and T. J. Sejnowski, "Learning and releaming in Boltzmann machines," *Parallel Distrilmted Processing*, vol. 1, pp. 45-76, 1986.
[23] J. Kovac, P. Peer, and F. Solina, "Human skin color clustering for face detection," in *Proc. Eurocon -International Conference on Computer As A Tool*, IEEE, 2003.
[24] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.
[25] R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696-706, 2002.
[26] J. Ngiam, Z. Chen, P. W. Koh, and A. Y. Ng, "Learning deep energy models," *in Proc. the 28th International Conference on Machine Learning (ICML-11)*, 2011.
[27] P. S. Nikkam and E. B. Reddy, "A key point selection shape technique for content based image retrieval system," *International Journal of Computer Vision and Image Processing (IJCVIP)*, vol. 6, no. 2, pp. 54-70, 2016.

[28] S. Karamizadeh, S. M. Abdullah, M. Zamani, J. Shayan, and P. Nooralishahi, "Face recognition via taxonomy of illumination normalization," in *Multimedia Forensics and Security*, Springer, 2017, pp. 139-160.
[29] S. Karamizadeha, S. Mabdullahb, E. Randjbaranc, and M. J. Rajabid, "A review on techniques of illumination in face recognition," *Technology*, vol. 3, no. 2, pp. 79-83, 2015.
[30] R. Lienhart and R. Hauke, "Filtering adult image content with topic models," in *Proc. 2009 IEEE International Conference on Multimedia and Expo (ICME 2009)*, 2009.
[31] C. Y. Liou, W. C. Cheng, J. W. Liou, and D. R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, pp. 84-96, 2014.
[32] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016.
[33] K. Nummiaro, E. Koller-Meier, and L. V. Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, no. 1, pp. 99-110, 2003.
[34] M. N. Patel and P. Tandel, "A survey on feature extraction techniques for shape based object recognition," *Image*, vol. 137, no. 6, 2016.
[35] C. Poultney, S. Chopra, and Y. L. Cun, "Efficient learning of sparse representations with an energy-based model," in *Proc. Advances in Neural Information Processing Systems*, 2006.
[36] Y. Raoui, E. H. Bouyakhf, M. Devy, and F. Regragui,. "Global and local image descriptors for content based image retrieval and object recognition," *Applied Mathematical Sciences*, vol. 5, no. 42, pp. 2109-2136, 2011.
[37] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. the 28th International Conference on Machine Learning (ICML-11)*, 2011.
[38] N. Sae-Bae, X. Sun, H. T. Sencar, and N. D. Memon, "Towards automatic detection of child pornography," in *Proc. 2014 IEEE International Conference on Image Processing (ICIP)*, 2014.
[39] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," *Artificial Intelligence and Statistics*, 2009.
[40] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85-117, 2015.
[41] J. Shayan, S. M. Abdullah, and S. Karamizadeh, "An overview of objectionable image detection," in *Proc. 2015 IEEE International Symposium on Technology Management and Emerging Technologies (ISTMET)*, 2015.

Sasan Karamizadeh received his M.Sc. and Ph.D. degrees in computer science all from University of Technology, Kuala Lumpur, Malysia, in 2012, and 2017 respectively. He is in post-doctoral degrees in Iran telecommunication research center. Image processing and face recognition are his interest's fields and has published several papers in international journals and conferences.

Abouzar Arabsorkhi received Ph.D. degrees at the University of Tehran in the field of information systems management. He is faculty member and director of the Network and System Security Assessment Unit in the Information and Communications Technology Research Institute. Over the past years, he has been involved in security management and planning, security architecture, risk management, security assessment and prototype certification, and the design and implementation of specialized security labs. The internet security of objects is one of his research interests. During the past 10 years, he has been taught in the field of information systems and e-commerce security.