SPC

International Journal of Engineering & Technology, 7 (2.33) (2018) 990-993

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Image anonymization using clustering with pixelization

Ria. Elin Thomas [1] *, Sharmila K. Banu [2], B. K. Tripathy [3]

[1] Masters Student, Scope, Vellore Institute of Technology, Vellore
[2] Assistant Professor, Scope, Vellore Institute of Technology, Vellore
[3] Senior Professor, Scope, Vellore Institute of Technology, Vellore
**Corresponding author E-mail: riaelinthomas@gmail.com*

Abstract

With the increasing usage of images to express opinions, feelings and one's self, on social media, and other websites, privacy concerns become an issue. The need to anonymize a person's face, or other aspects presented in an image for legal or personal reasons has sometimes been overlooked. Pixelization is a common technique that is used for anonymizing images. However, this technique has been proved to be a not-so-reliable technique, as the images can be restored using de-pixelization techniques. Clustering is usually used in relation to images, for image segmentation. When used in combination with pixelization, it proves to be an effective way to anonymize images. In this paper, the authors investigate the cons of using only pixelization, and prove how the use of clustering can improve the chances of anonymizing effec-tively.

Keywords: *Fuzzy C-Means Clustering; Image Anonymization; Pixelization; Privacy*.

1. Introduction

Social media is one of the platforms where millions of images are uploaded into the world-wide web on a daily basis. However, many of these contain images of people or things that may exploit privacy or may lead to legal issues. Pixelization and blurring are some of the techniques that have been used to avoid these issues. With the increase in sophisticated face detection, face recognition, and image restoration algorithms and technologies, the pixelization and blurring techniques have failed to achieve its purpose.

A framework [1] was defined which helps to assess the privacy protection solutions for video surveillance. Two face recognition algorithms were assessed, namely Linear Discriminant Analysis (LDA) and Principal Components Analysis (PCA). The CSU Face Identification Evaluation System was used to compare the two face recognition algorithms. These face recognition algorithms were applied to those images that were altered using privacy protection techniques such as pixelization, Gaussian blur, and scrambling. The authors concluded that pixelization and blurring was ineffective for the purpose of privacy protection. However, their observations were that the scrambling techniques were proved to be more efficient in comparison to the former two techniques. Another framework [2] was proposed that helps to protect privacy during crowd movement analysis. The face image data are first anonymized before sending it to the data center. In the data center, the authors have proposed to use the eigen face recognition, where the pattern matching is done in a low dimensional space. The anonymization is done by k-member clustering applied to the facial feature vectors. The authors concluded that although this did ensure privacy, the obtained knowledge was not as reliable, since anonymization led to information losses.

The authors in article [3] have investigated the applicability of Fuzzy clustering-based k-anonymization for the crowd movement analysis. Since, fuzzy clustering has the multi-cluster assignment feature, this helps to reduce information losses. The fuzzy clustering was applied to the eigen face features that would be sent to the data center. The authors concluded that with an appropriate fuzziness degree, the fuzzy clustering-based k-anonymization had advantage over the normal k-member clustering. In [4], the authors conducted a user study to discuss the effects of anonymization techniques that include different pixelization techniques. 103 users participated in this study, and they were made to verify whether people from the obscured images can be identified, and also whether their actions can be recognized.

A system [5] is designed to protect the privacy in the medical images. DICOM (Digital Imaging and Communications in Medicine) image information and the oncology patient records are anonymized here. Anonymization of the data is done by implementing policies based on the type of the user accessing the system.

Authors in [6] deal with the process of preserving the privacy of the data providers while combining their works in the database for datamining problems. M-privacy algorithm and multiparty computational protocol is used for anonymization instead of general k-anonymity and l-diversity for privacy preserving.

A machine-learning model [7] was proposed for anonymizing DICOM images. The model consisted of image preprocessing, classification algorithms, and de-identification. The authors were able to conclude that with RBM and Random Forest classifier, they were able to attain a 94% of precision, recall and F1-Score. Privacy protection filters like pixelization, Gaussian blur, blackener etc. were analyzed by a framework [8] that detects the presence of a filter, and classifies the type of filter that was used, along with the strength of the filter. An appropriate tool was then used to reverse the anonymization process which revealed that the filters were unable to achieve its purpose of protecting the identity of the people.

Authors [9] did a review on the different methods used to achieve image anonymization like blurring, pixelization, chaos cryptography etc., to verify whether these methods can be reversible. They evaluated these methods based on security and intelligibility. The methods were categorized under Transform-domain and pixel level, based on the method that is used by each of these techniques to anonymize the image. Authors in [10] performed an attack on the pixelization technique that was done on video streams. This was done by first taking the average of two frames, and then applying maximum-a-posteriori method to recover the image.

The paper is organized as follows: Section II explains the two algorithms that are used for anonymizing the image, i.e., pixelization and clustering. Section III shows how the algorithms were put together for implementing the anonymization algorithm, and Section IV shows the results that were obtained. Section V gives the conclusion to the paper.

2. Concepts

In this section, we provide some of the concepts to be used in the paper.

2.1. Pixelization

Pixelization is a strategy to make parts of a picture difficult to recognize by the human eye by misleadingly diminishing the picture resolution. It is achieved by splitting the image into M*M squares that are non-overlapping, where M is user-defined. The pixels within the image are replaced by the average value within each square.

$$Ip(x,y)=\frac{1}{n^2}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} I\left(\left\lfloor\frac{x}{n}\right\rfloor+i,\left\lfloor\frac{y}{n}\right\rfloor+j\right) \tag{1}$$

where x and y are the pixel coordinates and n is the block size.

Pixelization has been used for various purposes such as censorship and anonymization. It is a very common way to anonymize faces, number plates, gestures that are deemed unfit to show to the public through media, etc. However, through various de-pixelization techniques such as Bicubic Interpolation [11] and Cubic convolution interpolation [12], it has been proved that this technique cannot be relied for effective anonymization.

2.2 Fuzzy c-means clustering

Clustering has been defined as the process of putting more alike elements into groups such that elements in different groups are less alike than those in a single group. It has been noted that we the human beings study elements in the universe in clusters. This saves the reasoning time as if elements are in a single cluster then it becomes easier to study the characteristics of an element and extend it to those of the other elements in the cluster. Whatever knowledge is obtained for a single element or individual elements need not be repeated again and again [18].

The first clustering algorithm introduced is the hard C-means and the outputs of this algorithm are non-overlapping by nature. We find in real life that most of the cases we require the clusters to be overlapping; i.e. an element can belong to more than one cluster to certain degrees lying in the interval [0, 1]. The clustering techniques which generate such type clusterings are called uncertainty based clustering algorithms. There are several models of uncertainty available in the literature introduced so far. Each of these models depends upon one of the uncertainty based models like Fuzzy sets, Rough sets, Intuitionistic fuzzy sets or soft sets and their hybrid models like the fuzzy rough sets, rough fuzzy sets, intuitionistic fuzzy rough sets or rough intuitionistic fuzzy sets. The outputs in the case of these algorithms are fuzzy sets in case of Fuzzy C-means, rough sets in case of rough C-means and so on. The property of a fuzzy set is that we have graded membership values for the elements instead of crisp membership, i. e. an element either belongs to a cluster or not. The graded membership leads to partial belongingness of elements to the cluster. Similarly, the concept of intuitionistic fuzzy sets associates two functions with the set; one is called the membership function and the other one is called the non-membership function. In case of also these two two membership functions are in existence. But the non-membership function is just the one's complement of the membership function. So, it had no separate existence. However, in case of intuitionistic fuzzy sets, the sum of the membership values of any element lies in the interval [0, 1]. Hence in the case of intuitionistic fuzzy C-means, the clusters are such that the elements have both membership and non-membership values. This leads to an important notion called the hesitation function. This adds value to the uncertainty of belongingness of an element. The intuitionistic fuzzy c-means algorithm has been developed and studied in [19], [21].

Similarly, in case of rough C-means, the clusters are rough sets. The rough set notion depends upon the notion of uncertainty being captured by the boundary region of a set. However, the basic definition of a rough set depends upon an equivalence relation defined over the universe. This is because, the definition of knowledge introduced by Pawlak, the originator of the notion of rough set is that human knowledge depends upon the ability to classify objects in a domain. The classifications are disjoint subsets of the universe and when they are combined together by union we get the whole universe. The equivalence relations defined over a universe also decompose the universe into disjoint classes. It is easy to see that the two notions of classifications and the equivalence relations are interchangeable notions. So, for mathematical reasons Pawlak took equivalence relations to define rough sets. He introduced the notions of upper approximation and lower approximation with a set with respect to an equivalence relations which approximate the set from the lower and upper side by the way the set being included in the upper approximation and containing the lower approximation. Obviously, when the lower and upper approximation becomes identical we get a crisp set. Otherwise the difference between the upper approximation and the lower approximation is termed as the boundary of the set and is the region containing the uncertain elements.

It was observed in the beginning that the two models of fuzzy set and rough set are competing models and even people tried to establish the superiority of model over the other. But, it was established by two scientists Dubois and Prade in 1990 that far from being competitive they the models complement each other. Going a step ahead they combined these two models to propose the hybrid models of rough fuzzy and fuzzy rough models. It has been established since then that the hybrid models are more efficient than the individual components. It is worth noting that many such models have been proposed in the form of rough intuitionistic fuzzy sets and intuitionistic fuzzy rough sets and more importantly C-means algorithms have been proposed and studied for clustering data and have been applied to image segmentation [16, 17], Several applications of these algorithms can be found in some recent works [15, 20, 22],. However, in this paper, we focus our study by taking the fuzzy C-means algorithm only.

Definition 2.2.1: *A fuzzy set A defined over a universe W is determined by its membership function* m_A *such that* $m_A : W \to [0,1]$. *Thus for any element e in W,* $m_A(e)$ *assumes a value lying in [zero, 1].*

The notion of fuzzy set is an extension of the crisp set in the sense that every crisp set is associated with a function called its characteristic function. When m_A assumes values only zero or one, it reduces to a characteristic function and the corresponding fuzzy set reduces to a crisp set.

Definition 2.2.2: *An intuitionistic fuzzy set An over a universe W is determined by the membership and non-membership functions* m_A *and* n_A *such that* $m_A, n_A : W \to [0,1]$ *such that for any element 'e' from W,* $m_A(e)+n_A(e)\in[0,1]$.

So, the hesitation function h_A is such that for all e in W, $h_A(e)=1-\{m_A(e)+n_A(e)\}$.

When, $n_A(e)=1-m_A(e)$, the intuitionistic fuzzy set reduces to a fuzzy set. Here the hesitation function becomes a zero function.

Definition 2.2.3: *Let A be a set over a universe W and P be an equivalence relation over W. Let us denote the equivalence classes generated by P over W related to an element ‘e’ as* $[e]_P$.

Then we denote the lower approximation and upper approximation of A with respect to P by $L_P(A)$ and $U_P(A)$ respectively and define them as $L_P(A)=\{e \in W \mid [e]_P \subseteq A\}$ and $U_P(A)=\{e \in W \mid [e]_P \cap A \neq \phi\}$
When $L_P(A) \neq U_P(A)$, we say that A is rough with respect to P and it is said to be P-definable.
Fuzzy c-means clustering was first introduced by Dunn in 1973 [13]. Several authors have tried to improve the algorithm over the years since then. However, the algorithm used at present is the version enhanced by Bezdek in 1984 [14]. We would like to state that in the algorithm proposed by the concept of fuzzifier is used. He has used the fuzzifier to be a real number ‘m’, which has a range of values. In fact in his objective function Dunn had used 2 for the value of ‘m’. It was noted later that the value of ‘m’ lies in the interval [1.5, 2.5]. It has been taken that the ideal value of ‘m’ happens to be ‘2’. So, although Bezdek has extended the objective function of Dunn, the ideal case is the special objective function used by Dunn. The fuzzy c-means addresses the situations where a data can belong to two different clusters, and thus providing the “fuzzy” effect to the traditional k-means clustering algorithm. It follows on the minimization function of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} m_{ij}^{p}\left\|a_i - c_j\right\|^2 , 1 \leq p < \infty \tag{2}$$

where a_i is the ith data within the dataset, m_{ij} is the degree of membership of a_i in the cluster j, c_j is the centre of the cluster, ||*|| is any standard of measure to express the similarity between any data and the center of the cluster, and p is a real number greater than 1.
The membership function mij can be measured as follows:

$$m_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|a_i - c_j\|}{\|a_i - c_k\|}\right)^{\frac{2}{p-1}}} \tag{3}$$

The cluster center cj is measured as follows:

$$c_j = \frac{\sum_{i=1}^{N} m_{ij}^{p} . a_i}{\sum_{i=1}^{N} m_{ij}^{p}} \tag{4}$$

The whole process iterates until it meets the stopping criterion, i.e,

$$max_{ij}\left\{\left|m_{ij}^{(k+1)} - m_{ij}^{(k)}\right|\right\} < \epsilon \tag{5}$$

Where ϵ is between 0 and 1, and k is the number of iterations.
The algorithm is composed of the following steps:

1) Initialize M = [mij] matrix, M(0)
2) At kth step: calculate the centers vectors C(k) = [cj] with M(k)

$$c_j = \frac{\sum_{i=1}^{N} m_{ij}^{p} . a_i}{\sum_{i=1}^{N} m_{ij}^{p}}$$

3) Update M(k), M(k+1)

$$m_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|a_i - c_j\|}{\|a_i - c_k\|}\right)^{\frac{2}{p-1}}}$$

4) If || M(k+1) – M(k)|| < ε the STOP, otherwise return to step 2

3. Implementation

Python packages were used to implement the anonymization. The Pillow packages were used to resize the image and the Nearest Neighbor Interpolation sampling filter was applied. The Geotiff package was used to read the image and the gdal package converts it into grayscale image by taking only a layer of the raster band. The Pillow package was used to convert it into a numpy array, which is then used to give as input to the fuzzy c-means clustering module.

4. Results

In this section we take a specific image of a human being in a specific position and generate the pixelized image from it then we apply clustering to generate the segmented mage.

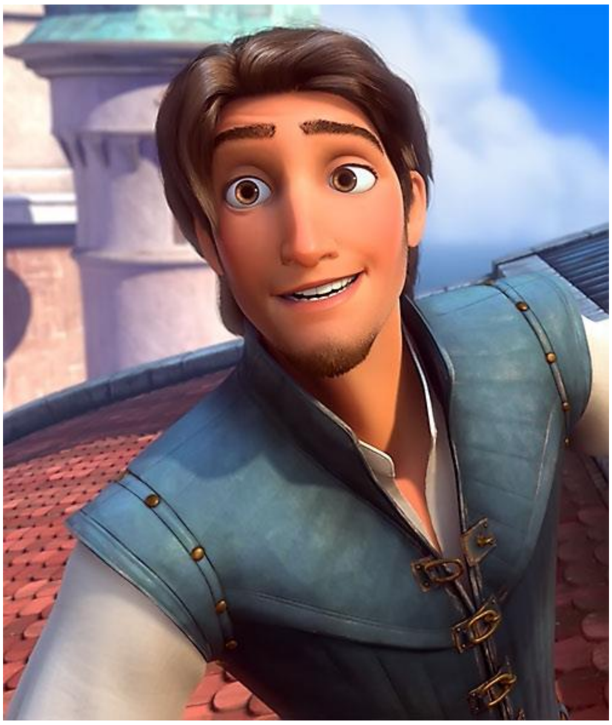

Fig. 1: Original Image.

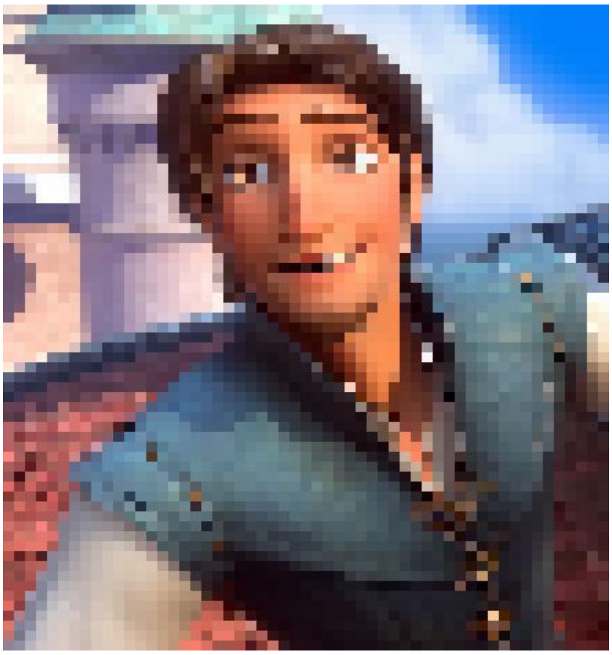

Fig. 2: Pixelized Image

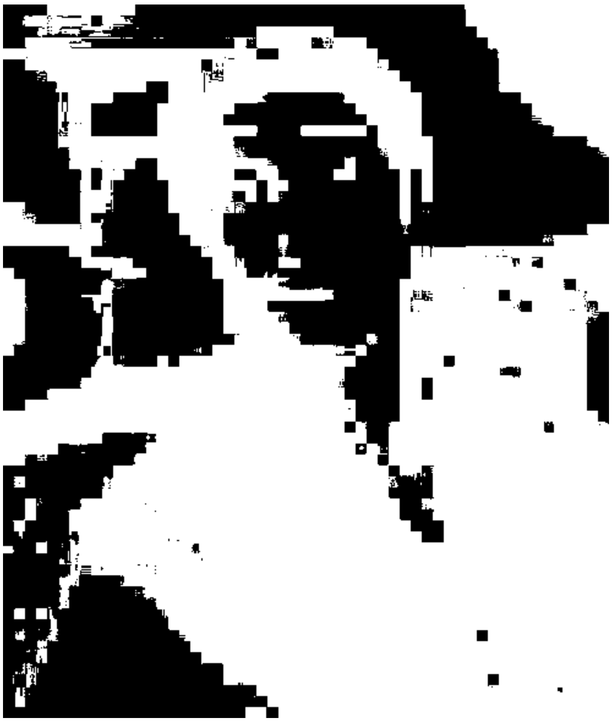

Fig. 3: Clustered Image.

The original image (Fig 1) is pixelized (Fig 2), and is then clustered (Fig 3) using fuzzy c-means clustering to obscure the image against face recognition algorithms. The obscured image is now more anonymized than when pixelized.

5. Conclusion

With pixelization being used as a common method to obscure images, the authors have proposed to use clustering algorithm along with pixelization to better secure it against de-pixelization methods and other possible techniques that may give away the identities of the people in a picture. Since the clustering algorithm uses complex numerical expressions, it becomes almost impossible to de-cluster the image, thus providing more security for maintaining the privacy of the image. This can be extended onto videos, like surveillance, where certain identities need to be anonymized, before publishing it.

References

[1] Dufaux, F., & Ebrahimi, T. (2010, July). A framework for the validation of privacy protection solutions in video surveillance. In Multimedia and Expo (ICME), 2010 IEEE International Conference on (pp. 66-71). IEEE.

[2] Honda, K., Omori, M., Ubukata, S., & Notsu, A. (2015, June). A privacy-preserving crowd movement analysis by k-member clustering of face images. In Informatics, Electronics & Vision (ICIEV), 2015 International Conference on (pp. 1-5). IEEE.

[3] Honda, K., Omori, M., Ubukata, S., & Notsu, A. (2015, November). A study on fuzzy clustering-based k-anonymization for privacy preserving crowd movement analysis with face recognition. In Soft Computing and Pattern Recognition (SoCPaR), 2015 7th International Conference of (pp. 37-41). IEEE.

[4] Birnstill, P., Ren, D., & Beyerer, J. (2015, August). A user study on anonymization techniques for smart video surveillance. In Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on (pp. 1-6). IEEE.

[5] Shahbaz, S., Mahmood, A., & Anwar, Z. (2013, December). SOAD: Securing oncology EMR by anonymizing DICOM images. In Frontiers of Information Technology (FIT), 2013 11th International Conference on (pp. 125-130). IEEE.

[6] Indhumathi, R., & Priya, S. M. (2014). Data Preserving By Anonymization Techniques for Collaborative Data Publishing. International Journal of Innovative Research in Science, Engineering and Technology, 3(1), 358–363.

[7] Monteiro, E., Costa, C., & Oliveira, J. L. (2015, August). A machine learning methodology for medical imaging anonymization. In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE (pp. 1381-1384). IEEE.

[8] Ruchaud, N., & Dugelay, J. L. (2016). Automatic Face Anonymization in Visual Data: Are we really well protected?. Electronic Imaging, 2016(15), 1-7.

[9] Pantoja, C., Arguedas, V. F., & Izquierdo, E. Anonymization and De-identification of Personal Surveillance Visual Information: A Review.

[10] Cavedon, L., Foschini, L., & Vigna, G. (2011, August). Getting the Face behind the Squares: Reconstructing Pixelized Video Streams. In WOOT (pp. 37-45).

[11] Keys, R. (1981). Cubic convolution interpolation for digital image processing. IEEE transactions on acoustics, speech, and signal processing, 29(6), 1153-1160.

[12] Dong, W., Zhang, L., Shi, G., & Wu, X. (2011). Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. IEEE Transactions on Image Processing, 20(7), 1838-1857.

[13] Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

[14] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.

[15] B.K.Tripathy and P.Swarnalatha: A Comparative Study of RIFCM with Other Related Algorithms from Their Suitability in Analysis of Satellite Images using Other Supporting Techniques, Kybernetes, vol.43, no.1,(2014), pp. 53-81

[16] B.K.Tripathy, R. Bhargav, A. Tripathy, E. Verma, Raj Kumar and P.Swarnalatha: Rough Intuitionistic Fuzzy C-Means Algorithm and a Comparative Analysis, COMPUTE"13, Aug 22-24, Vellore, Tamil Nadu, India Copyright 2013 ACM 978-1-4503-2545-5/13/08

[17] B.K.Tripathy and R. Bhargav: Kernel Based Rough-Fuzzy C-Means, PReMI, ISI Calcutta, December, LNCS 8251, (2013), pp.148-157

[18] Swarnalatha P, Tripathy B.K., Nithin, P. L. and D. Ghosh: Cluster Analysis Using Hybrid Soft Computing Techniques, CNC-2014International Conference of Network and Power Engineering ,Proceedings of Fifth CNC-2014,pp. 516-524

[19] B.K.Tripathy, Avik Basu and Sahil Govel: Image segmentation using spatial intuitionistic fuzzy C-means clustering, proceedings of the IEEE ICCIC2014, (2014), pp.878-882

[20] B.K.Tripathy and D. Mittal: Efficiency Analysis of Kernel Functions in Uncertainty Based C-Means Algorithms, International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015, Article number 7275709, pp. 807-813 (2015).

[21] B.K. Tripathy, Deepthi P.H. and Dishant Mittal: Hadoop with Intuitionistic Fuzzy C-means for clustering in Big Data, Advances in Intelligent Systems and Computing, Volume 438, 2016, Pages 599-610.

[22] B. K. Tripathy, Akarsh Goyal and Rahul Chowdhury: MMeNR: Neighborhood Rough Set Theory Based Algorithm for Clustering Heterogeneous Data, International Conference on Inventive Communication and Computational Technologies (ICICCT 2017), (2017), pp.323-328.