International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET231027

Icensor : Unwanted Image Detection and Censoring

Dr. Manju Bargavi[1], Sakshi Dhruva[2], Tenzin Kunsang[3], S Subham Patra[4], Tenzin Nyima[5]

[1]Professor, Department of CS & IT, Jain University, Bangalore, Karnataka, India

[2,3,4] MCA, Department of CS & IT, Jain University, Bangalore, Karnataka, India

[5]School of Commerce, Jain University, Bangalore, Karnataka, India

ARTICLEINFO

Article History:

Accepted: 05 March 2023
Published: 20 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
75-85

ABSTRACT

The globe today, practically everyone uses the internet, which is filled with a vast amount of information and content, the majority of which includes pornographic and violent photos and movies. Social media has grown in importance in today's society as a result of the expansion of the internet. This has increased the risk of privacy invasion, which includes the release of private photos that should not be shared because they violate the privacy of some people. Today, even a young child can easily access these materials. Recent image leaks from prominent social media applications and the use of private photos by clever algorithms have caused the public to re-evaluate the need for individual privacy when uploading images on social media. The process of sharing photos on social networking sites is complex in and of itself, and the measures in place to safeguard privacy in daily life are labor-intensive and fall short of providing tailored, precise, and adaptable privacy protection. We have found that techniques like "privacy intelligence" solutions, which concentrate on current privacy issues related to online social networking image sharing, are effective. An optical character recognition-based system that filters photographs with sensitive text, in addition, a visual algorithm that filters out pictures that aesthetically resemble those on an image blacklist is used. These measures can help stop the spread of any delicate content on social media.

Keywords: Unwanted Image Detection, Nude Image Detection, Image Censoring

I. INTRODUCTION

The easy availability of the internet has both positive and negative effects on individuals and society. The easily available internet makes audiences search for any content and poses many advantages but there are audiences that are very sensitive about the violence that they are accidentally exposed to such type of violent content that they might get disturbed. When we talk about illicit content on social media it can be

in many forms. Illicit images on social media refer to any visual material that is illegal, harmful, or offensive that is shared or distributed through social media platforms. Such images can include Child sexual abuse material that includes images or videos that depict the sexual abuse or exploitation of minors, Revenge porn: non-consensual sharing of explicit images or videos of an individual, typically with the intent to cause harm or embarrassment, Pornographic images involving animals: images or videos that depict sexual acts involving animals, Images promoting terrorism: propaganda or recruitment material for terrorist groups. Images promoting hate speech: visual material that promotes hatred or discrimination based on race, religion, gender, sexual orientation, etc.

So Icensor is a software application that one can install on their computer that can automatically detect and censor these contents by saving young minds. After the installation of the Icensor, the admin can access the Icensor control panel and manage the censoring system whether to keep it on or off. The audiences/children can now able to browse the internet without any disturbing content. The contents are all censored by the Icensor. The system is password protected so the user apart from the admin can't make the changes. The admin has the right to exclude some websites that he/she doesn't want to be censored and can turn off the censoring when not required. Icensor is primarily concerned with masking obscene photographs on the internet. The intended significance of the image is hidden through censorship. Icensor is superior to the current methods that completely obscure the contents and defeat their intended purpose.

The project is divided into 3 sections:

I.I Extraction of Images

A customized data extraction tool (web scraper) is used for extracting an image from web pages. To make sure that websites from across the Internet can be found in search engine results, a web crawler is employed to index the content of such websites. The purpose of web scrapper is to accurately and quickly extract images from the web. Web scrapper collects only images here. The First step is extracting the data from the web so a scrapper is used for extraction.

I.II Detection of Objects

The object is detected using Mask RCNN. RCNN is a Region-based Convolutional Neural Network to solve the instance segmentation problem in ML [9]. RCNN separates different objects from the content. To train this model, an image of a gun is given.

I.III Masking

A mask is attached to the object so that the image can be censored. the censored image is then sent back to the browser.

II. RELATED WORK

Content censoring can be implemented client-side (i.e., on the application itself) or server-side (i.e., on a remote server). In a client-side implementation, the program running on a user's device contains the censoring rules to be applied. A built-in list of keywords is frequently present in applications that provide client-side keyword filtering [6][7][8] and can be updated whenever a client connects to a server [1]. Prior to being sent, a user's message is not delivered if it contains any prohibited keywords. The censorship rules in a server-side implementation are located on a distant server. When a message is transmitted, the server examines this for the presence of prohibited keywords and, if any are found, stops the message.

Reverse engineering techniques have been used in the past to obtain keyword lists that are employed to turn on censorship on messaging apps, live streaming services, and online games. Sample testing is used in several server-side censorship experiments, in which researchers either assemble material in a collection, they believe would remain prohibited through a platform, provide the sample there, and record the

results, or they choose public posts and keep an eye out for deletion.

WeChat, one of China's popular social networking applications which has more than one billion monthly active users, is the most widely used social media network in China [3]. Group and private messaging, Moments (a feature for uploading messages and images similar to Facebook's Timeline), and other social networking services capabilities are all included in its functionality. We discovered that Tencent automatically and in real-time filters conversation photographs on WeChat based on the language they include as well as how visually similar they are to images on a blacklist [1]. Tencent makes immediate filtering possible by maintaining a hash index comprising the MD5 hashes of the pictures users of chat platforms send [4]. An image is not filtered if its MD5 hash, which is supplied not a part of the hash index over the chat platform. Instead, it's put in a queue for computerized analysis. When an image with the same MD5 hash is sent later, it will be censored if it is found to be sensitive since its hash will be added to the hash index.

The vulnerability of neural network-based image classifiers to adversarial instances, or images containing subtle changes that lead to incorrect classification, is well recognized. Recent research has demonstrated the ability to generate adversarial examples by estimating the network's gradients, in contrast to earlier work that relied on a Whitebox threat model with known implementation information, including the target network's training gradients [2].

However, the study assuming the strictest threat model still presupposes that the attacker may obtain probability scores for the top k categories for any given image, which limits its applicability. WeChat does not use machine learning to filter images, and even if it did, our analysis demonstrates that the threat model for evading a censorship filter is much more constrained than what is currently assumed in the literature on adversarial situations because the only signal provided is whether an uploaded image is filtered. Our approach differs from that of the adversarial examples literature in that we create evasion techniques in identifying then taking advantage of additional crucial Whether or not the filter employs machine learning categorization, implementation specifics of the filtering algorithm are still helpful.

A common image recognition technique known as Content-Based Image Retrieval (CBIR), image retrieval is a computer system that is used to browse, search for, and retrieve images from a sizable collection of digital images. In order to execute retrieval over the words annotated to the photos, the majority of conventional and traditional techniques of image retrieval employ some means of adding information, such as captioning, keywords, titles, or descriptions to the images. Automatic picture annotation has been the subject of much research since manual image annotation is time-consuming, costly, and difficult. A number of web-based picture annotation tools have also been created as a result of the growth of social web apps and the semantic web. CBIR (Content-Based Image Retrieval) evaluates the color and geometry of pictures to determine how well they work. With this method, the backdrop, texture, and form are basic visual elements. Based on the skin content of an image that is more than a certain threshold, the data is then used to create results for queries [13]. Despite the fact that CBIR has been the subject of much research, nothing has been done to completely develop an engine that is just focused on the search of picture content. The majority of CBIR systems struggle to integrate high-level elements like semantic information with low-level picture attributes like intensity, color, texture, shape, and spatial limitations. A CBIR system has been proposed to deal with this issue that uses a shape thesaurus, instead of a text thesaurus, to link the low-level features extracted with semantics Shape thesaurus may inject more features into CBIR-related approaches than a text thesaurus. Thus, the system

offers better efficiency by combining shape information with the text thesaurus [14].

A chain code principle method to represent shape descriptors has also been used in the field of image recognition. New definitions can be derived from existing ones in this manner. The chain code is used to generate the shape thesaurus because the compared descriptors are straightforward to employ with other methods, such as Fourier Shape Descriptors (FSD) and Moment Invariants (MI). We use FSD, MI, Finite Element method (FEM), Turning Function, and Wavelet Descriptor to obtain image shape to compare with other images' shape thesaurus [15].

A digital feature retrieval approach for JPEG picture files was presented. The approach extracts JPEG byte stream data at the packet level and searches for suspicious files using its Direct Current (DC) coefficients. As baseline JPEG files employ Variable Length Codeword (VLC) encoding, different DCT coefficients use 21 varying lengths of codes. Each Minimum Coded Unit is responsible for the sequential extraction of Y, Cb, and Cr. End of Block (EOB) values enable the extraction of the next block's DC components. For comparisons, DC components extracted from JPEG packets with the same feature database are employed [16].

Dynamic filtering is a proactive approach to limiting material depending on predefined criteria like textual content and website address. This filtering strategy is more complicated than static filtering in terms of blocking access to improper items. It employs a comprehensive dictionary-based text analysis to decide whether or not the information is acceptable for viewing. Static filtering employs a predefined blacklist approach to restrict information by utilizing certain keywords inside web pages or their URLs, which may also be banned independently.

Host-based filtering allows users to install a program that integrates with the operating system to give protection, or software that integrates with a primary Internet access application, such as a web browser. NetNanny and Cyber Patrol are two commercially accessible examples of this sort of filtering. In most cases, proxy-based filtering is used in enterprises to block unwanted information and websites. This filtering method can be utilized within a server relay used by employee workstations to give internal Internet access. Workstations connect to this server, which filters previously banned websites by addresses or keywords and blocks access to these sites and their content. Because of load issues, this form of filtering often uses a static technique.

Support vector Machine learning model for detection of porn content by exploiting rudimentary information from pornography and utilizing this knowledge to determine whether or not a particular photo belongs to pornography. To begin, the skin region from pictures is extracted and investigated on the association between the skin region and the non-skin region, Skin is one of the most significant elements to extract when identifying human images. This aspect is especially crucial in pornography since pornography contains a considerable amount of the skin region. Skin pixel identification can be as simple as claiming that an RGB (Red, Green, Blue) pixel is skin when R > G or R > B, or perhaps both. We can extract additional color attributes from skin pixels after identifying them in skin filtering. We may use the color distribution of skin pixels in color space to apply a pattern-matching algorithm over a smaller region with little patterns from an obscene database. The obscene database contains some common but distinct pornographic skin pixel distribution. Also, we may add several "back-areas" that frequently feature in pornography to the database. If there is a match, it suggests that the photo is more likely to be pornography, and we may use it to determine pornography. The second phase that is skin pixel correlation is the collection of relationships that exist between skin areas and their non-skin neighbors. These surrounding locations can occasionally be beneficial in identifying obscenity. Hair is one of the

valuable indicators that may be utilized to identify pornography. Moreover, this contributes another characteristic known as hair-inside-body. The skin filtering method may separate a photo into two parts: the background and the skin region. We may improve SVM prediction by determining the fraction of "skin pixels" in such skin areas, as well as the smoothing and color correlation of skin regions. To summarize, the six criteria used to assess whether a photo is pornographic or not are skin percentage, pornography-weight, skin area geometric distribution, skin pixels in skin correlation, hair-inside-body, and skin-region-smoothness. The correlations are then fed into a support vector machine (SVM), a powerful classification tool with learning capabilities. The data that is fed into the SVM train includes the six characteristics as well as a label that indicates whether the photos are pornographic or not. Lastly, we re-enter this data for cross-validation, which allows us to fine-tune prediction accuracy. After a period of training, the model reached around 75% accuracy, a 35% false alarm rate, and just a 14% miss-detection rate [17].

The bag of visual words model is inspired by the bag of words model, in which each picture is defined by an unsorted set of discrete visual words obtained by the discretization of local descriptors. In this BOVW approach illegal picture, identification is used, in which images are represented as a histogram of visual words. The visual words represent local characteristics retrieved from photos, and the vocabulary is task-specifically learned from a training database. As local features, we extract image patches around difference-of-Gaussian interest points that are scaled to a common size and then PCA (Principal Component Analysis) transformed leaving 30 coefficients to reduce dimensionality. We use the training technique for unsupervised training of Gaussian mixture models to build a visual lexicon. Starting with a single Gaussian, this technique iteratively divides each existing density in the direction of its variance to produce a collection of 2#splits densities. The taught visual vocabulary may recognize commonly recurring patterns in the training data. The BOVW model can collect almost enough color information that an additional skin color model is not required [19].

III. METHODS

III.I OCR-Based Filtering

Different methods may be employed by OCR algorithms to recognize text. It is true that other algorithms and WeChat's OCR algorithm exchange implementation details, nevertheless, at a high level. The first thing that many OCR algorithms do is transform a color image to grayscale, which reduces it to only black, white, and light grey because many OCR algorithms do not function directly on color images. By doing this, they greatly simplify text recognition because they only need to work on one channel [1]. If the submitted color photographs were converted to grayscale by the OCR system, we made test images that would avoid filtering in order to evaluate if WeChat's OCR filtering method converts color images to grayscale [1]. In order to make the text invisible to the OCR system we made the images with text that is submerged in the color of an image in such a way that it is clearly readable by someone reading it in color after it has been converted to grayscale. The OCR algorithm turns images to grayscale if they managed to avoid being censored.

In theory, any function of the red, green, and blue intensities of a color pixel might be used to determine that pixel's grey intensity. After conducting this experiment, it was discovered that the only way to consistently avoid filtering for each of the tested colors was to select the level of grey background intensity suggested by the luminosity calculation. With most hues, the other formulas were unable to avoid censoring. Only when using red or cyan text did the averaging method avoid filtering, and only when using green or magenta text did the lightness algorithm avoid filtering.

III.II Collaborative Filtering

Collaborative filtering systems sort information according to ratings from other users who share the same interests as the user as well as the user's prior preferences. It is frequently used, particularly in e-commerce applications, numerous filtering systems, and recommender systems. Examples of such systems include Amazon.com and e-Bay, which analyze a user's prior buying behavior to suggest new things [12].

The calculation of similarity between user interests is a component of collaborative filtering systems. Using many tools, including the Pearson correlation coefficient, the similarity between users' interests is assessed. The algorithm determines the degree to which users' assessments of the same item are similar by gathering feedback directly from them or inferring it from their behavior. Ratings may be expressed explicitly on a scale of one to ten, or they may be implied by actions like mouse movements, clicks, and purchases. The users are then organized into groups based on the determined similarity measures, and moving forward, the user is advised to purchase certain things depending on the advice of other group members [12].

There are two types of collaborative filtering: One is model-based, while the other is memory-based. Whereas model-based collaborative systems utilize models to promote products by estimating models based on ratings, memory-based collaborative filtering systems use neighborhood principles. Several machine learning algorithms, including Bayesian networks, clustering, and Markov models, are used to create the models [12].

III.III Blob Merging

Most OCR systems next use a thresholding approach to convert each pixel in the grayscale image, which may originally be some color of grey, to either wholly black or completely white, leaving no shades of grey in between. This is done after converting a colored image to a grayscale [1]. The next phase in certain OCR algorithms is blob merging, which comes after thresholding. These algorithms attempt to identify which blobs in an imaging correlate to each character in order to identify each character. Many symbols, like the English letter I am composed of separate parts.

III.IV Visual-Based Filtering

We discovered that WeChat uses a second filtering technique the restriction of images without the text in addition to OCR-based filtering [1]. This filtering algorithm compares the visual resemblance between a given image and those on a prohibited image list. To evaluate alternative hypotheses regarding how the filter operated and to offer instructions for avoiding the visual-based filter, WeChat made changes to politically sensitive photographs that WeChat had blacklisted. To limit the analysis despite learning that photographs filtered using visual-based approaches were usually removed so rapidly that they were never available in the other account's view, they once more only took into account images that were censored during 60 seconds of being posted.

They found that the visual-based implement is quicker than the OCR-based one, suggesting that their method is computationally cheaper than the one employed for OCR filtering.

III.V Gray Scale conversion

Testing of the visual-based algorithm was done to see if and how it converts photos to grayscale, like how the OCR-based algorithm was tested. For testing the visual-based algorithm, as opposed to using an OCR-based technique, where text was in the forefront, the foreground was made up of White pixels from a monochrome image. WeChat threshold is an editorial cartoon with the flags of Hong Kong and the PRC in black and white [1]. They used the image's white pixels as the foreground and its black pixels as the background after making sure the image was still filtered even after being threshold. They once more chose the background according to three different grayscale algorithms for each of the six foreground

colors. As with the OCR algorithm test, it was discovered that the luminosity formula [10] was the only one that consistently filtered the photos.

III.VI Edge Detection

Edge detection methods used in image processing based on different operations are frequently utilized. Although it is sensitive to noise, it could be able to detect the variation in grey levels. In image processing, edge detection is a crucial process [5]. It is a crucial tool for scene analysis, image segmentation, and pattern recognition. An edge detector is a high-pass filter that may be utilized to extract the edge points in a picture. An image's edge is a delineation when there is a sharp change in brightness. In image processing, an edge is commonly considered one type of singularity. Simply put, singularities in a function are discontinuities where the gradient approaches infinity. Despite the fact that picture data is discrete, local gradient maxima are typically used to define an image's edges.

There are a lot of edges between backgrounds, objects, and backgrounds, as well as between primitives and primitives. The gray's discontinuity reflects the edge of an object. In order to identify an edge, one must first analyze how an individual image pixel changes in a grey area.

Next, one must use the variation of the edge's neighboring first or second order to identify the edge. The local operator edge detection approach is the name given to this method. Edge extraction is therefore a crucial approach in feature extraction and graphics processing. Edge detection's fundamental concept is as monitors: To start, use the edge enrichment worker to draw attention to the image's immediate edge. Next, establish the cutoff for obtaining the edge point set and determine the pixel "edge strength." However, possibly not be a continuous edge that was observed due to the noise and blurry image.

III.VII Canny Operator Edge Detection Technique

The Canny edge detector, an edge detection instrument, uses a multi-stage approach to recognize various edges in images. John F did so in 1986. Canny gave it life. In addition, Canny created an edge detection computational theory that clarifies how and why the technique works. The Canny filter is a multistage edge detector. A filter with a Gaussian derivative is used to determine the gradients' intensities. The Gaussian makes the noise in the image less obvious. Then, by removing pixels with insufficient gradient magnitude, potential edges are condensed to 1-pixel curves. Edge pixels are either kept or discarded using hysteresis thresholding that is performed to the gradient magnitude [5].

The Canny contains three variables that can be changed: the Gaussian's width (the wider the Gaussian, the messier the image), as well as the low and high thresholds for hysteresis thresholding. The basic standards intended for edge recognition are as follows - Edge detection with low error rates should accurately recognize all of the edges that are discernible in the image. The operator should be able to pinpoint the center of the edge with their edge point detection. Picture noise should, if at all possible, prevent the creation of fake edges, and an image’s edge should only be indicated once [5].

In the figure given below fig., 1 shows the working of the algorithm: Smooth the images using a Gaussian filter to remove noise. Recognize the image's intensity gradients. Use non-maximum suppression to eliminate erroneous edge detection responses. Use a twofold threshold to identify possible edges. Track the edge by hysteresis: Complete the edge detection by suppressing all additional weak and unconnected edges.

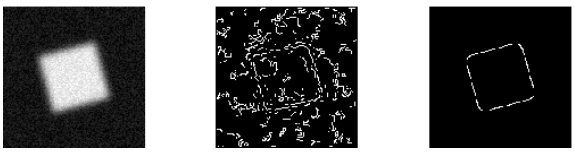

Fig 1: noisy image

III.VIII Perceptual Hashing

A technique for creating a hash from an image is called perceptual hashing which allows for efficient comparison of similar images by assigning the same or similar hashes to them. Many social media platforms, including Facebook, Microsoft, Twitter, and YouTube, employ it to filter out illicit content [1]. It is possible to create a hash that exhibits translational invariance using spectral techniques. The widely used open-source implementation pHash uses the non-translationally consistent discrete cosine transformation, to calculate a hash.

Images are made up of a group of pixels. Each pixel is made up of three values that represent the fundamental red, green, and blue hues (RGB). Each 8-bit pixel in a typical digital picture has a value between 0 and 255. The majority of perceptual hashes, often have three following steps [11]:

1. First, reduce the size of the image and convert it to grayscale.

2. Calculating the average grayscale pixel value.

3. For each pixel, encode 1 if the color is lighter than the average, otherwise 0.

As there are several options for each of these stages, the five required features of a hash are sometimes compromised. These compromises are between distinctiveness and resilience, and the specific manipulations to which the hash is resilient. For instance, a hard hash is very distinct yet shows little resistance to even a minor content change. On the other hand, a global hash with a threshold of =255 that is based on the average per-channel pixel color is extremely robust but completely non-distinct. Similar to how certain hashes can withstand changes in brightness and color but not necessarily in rotation and cropping [11].

There are many factors that are based on the choices of hash that includes [1]:

1. Scale: When working with the size of a big social network where millions of contents are uploaded daily, we require a highly efficient and distinct hash. Even a 1/100 or 1/10,000 false positive rate (mapping two photos wrongly) is unjustified at this size.

2. Tolerance: Resilience may not be as crucial when attempting to restrict the upload of images depicting child sexual abuse as it is when attempting to restrict the upload of, say, legitimate adult pornography.

3. Security: Non-reversibility may be less crucial when attempting to prevent the upload of copyright-infringing content than in other, more sensitive sectors.

III.IX Image Resizing

This subject discusses how altering the dimensionality of pictures impacts the program's ability to recognize them. For example, the authors [1] discovered that WeChat's filter could clearly screen sensitive photos regardless of scale as long as the aspect ratio was retained. We wanted to see if and how WeChat normalizes the dimensions of uploaded photographs to a canonical size. They chose to investigate five distinct hypotheses to get answers to these questions: (1) Pictures are scaled proportionately such that their width is a certain value, such as 100. (2) Pictures are adjusted proportionately such that their height equals a certain amount, such as 100. (3) Pictures are adjusted proportionately such that their maximum dimension is some number such as 100. (4) Pictures are downsized proportionately such that their lowest dimension is a number such as 100. Both dimensions are scaled to a given size and proportion, such as 100100, based on two factors.

If the last hypothesis is right, then WeChat's filter should be resistant to changes in a sensitive image's aspect ratio, because any aspect ratio changes would be eliminated when the image is resized to a fixed aspect ratio. Then they stretched the fifteen photos to test this idea. Each image was stretched 30% thinner and 30% shorter. Where they discovered that extending the

photographs was quite efficient at avoiding the filter, as all of the images extended shorter, as well as all of the thinner images save for a doodle, escaping filtering. This implies that the previous theory is erroneous [1].

To put hypotheses 1–4 to the test, they produced the following predictions: (1) If photos are downsized proportionately depending on their width, adding extra space to their width will avoid filtering, but adding it to their height will not. (2) If photos are downsized proportionately depending on their height, adding extra space to their height will prevent filtering. (3) If photos are downsized proportionately based on their biggest dimension, adding extra space to that dimension will prevent filtering. (4) If photos are downsized proportionately depending on their lowest dimension, adding extra space to that dimension will prevent filtering [1].

To put these predictions to the test, they picked ten filtered photos, five of which had a height that is no more than 2/3 of their width (we call these wide images) and five of which had a height that is no more than 2/3 of their width (we name these tall images). They then altered each image by adding blank black space the size of 50% of its width to its left and right sides, as well as black space the size of 50% of its height to its top and bottom sides. Repeat these steps, but this time use 200% of the corresponding dimensions [1].

They discovered that wide photographs with more space added to their width and tall images with additional space added to their height were always filtered. This is consistent with hypothesis 4, which states that WeChat resizes submitted images based on their smallest dimension because it predicts that adding space in this case will not affect the scale of the original picture contents once the image is resized. They also discovered that 4 out of 5 wide photos with space added to their height and 3 out of 5 tall images with space added to their width escaped screening, implying that the uploaded image was downscaled more than the matching one on the blacklist [1].

With only one fewer tall image filtered, the results between adding 50% and 200% more space were quite consistent. This consistency is to be expected because the smallest dimension hypothesis states that adding more space after the picture has already become square has no effect on its scaling. It's unclear why certain photographs were still filtered—two tall images with additional width and one wide image with extra height. It's probable that WeChat's filtering mechanism is resistant to scale adjustments. Nevertheless, versions of these photos with more space or another border or information added in these spaces may also be on the blacklist. A reverse Google Image search revealed that several photos with identical spacing had previously circulated on the Internet. Adding height to wide photos or width to tall images, on the other hand, was often a successful approach for avoiding filtering while retaining the image's aesthetic appeal [1].

III.X Sliding Window

Sliding windows are essential in object categorization because they allow us to pinpoint exactly where an object is in a picture. A sliding window is a rectangular section with a definite width and height that "slides" over a picture. In this section the authors [1] are concerned with whether the algorithm is not simply translationally invariant but whether it can find an image inside of another image regardless of its surrounding contents.

They discovered that WeChat's server-side image compression increased compression for bigger photos and that the ensuing compression artifacts might lead submitted images to avoid screening during testing. The experiment was meticulously constructed to account for this. Taking five wide and five tall photographs, by increasing their canvases by each side on their largest dimensions. They began by extending the canvases with darkness. Because many image procedures, such as thresholding and edge detection, are sensitive to the distribution of pixel intensities in an image, they additionally expanded the canvases

with multiple copies of the picture itself to adjust for this. To accommodate for WeChat's compression, they downloaded every picture they made and cropped off the enlarged canvas, returning it to its original size if it evaded screening. If this image still avoids filtering when uploaded, they determined that it is due to extra compression artifacts rather than the contents of the enlarged canvas [1].

They also discovered that once a sufficiently high number of duplicates were added, pictures expanded with their own duplicates avoiding filtering, and none of these evasions could be explained by image compression. In contrast, pictures expanded with blank canvases were either filtered or their evasion could be explained by image compression in all but one test. These findings imply that, even when we add extra content to an uploaded picture in such a way that the distribution of pixel intensities remains constant, these contents influence WeChat's capacity to classify the uploaded image as sensitive [1]. It shows that WeChat may not compare photos using a sliding window technique that ignores items outside of the window. Instead, it appears that the photos are assessed as a whole and that adding intricate patterns outside of a banned image's initial canvas can circumvent filtering [1].

III.XI Image Matching

Image matching is another valuable host-based use of our method that is employed to match black-listed images in the network. This approach does not rely on previously encountered imaging data and may be used to modify pictures. This overcomes the limitation of conventional hash-based matching algorithms, which rely on consistent data for correctness. When any modifications in a given picture (even one pixel) are provided, hash methods are prone to failure, resulting in a mismatch. Illegal photos on the internet are frequently saved and shown as thumbnails, which are then enlarged, watermarked, cropped, and manipulated in a variety of ways [18].

To do image matching, feature vectors from target photos are retrieved and saved in a database using MLE or SLWE. Every picture that passes through the system is first computed by its feature vector using Maximum Likelihood Estimator (MLE) or Stochastic Learning Weak Estimator (SLWE), depending on which database is utilized, and then compared to a database. If the distance is less than a certain threshold, the outcome of a match is recorded. If the threshold is not met, the result is reported as a non-match [18].

IV. CONCLUSION

The scope of this study is to safeguard people's privacy and prevent any negative events from happening on social media in the form of text or images that are irrelevant to be posted or shared further so that does not lead to something bad in the future.

This Icensor efficiently assists in the identification and detection of hazardous items or undesirable pictures. Modern materials are required to support particular applications rather than transducers for many cutting-edge sensor technologies, such as photon-scattering and laser acoustic technologies.

Through a lot of research, we came across some of the methods and concepts that help to filter out illicit content. When a message is sent, the server examines it for the presence of prohibited keywords and, if any are found, stops the message.

This survey paper can be used in client-side or server-side applications to censor or prevent things from happening on social media applications.

V REFERENCES

[1] Knockel, J., Ruan, L., & Crete-Nishihata, M. (2018, August). An analysis of automatic image filtering on WeChat Moments. In *FOCI@ USENIX Security Symposium*.

[2] Ch i Liu, T. Z. (2020, August). Privacy Intelligence: A Survey on Image Sharing on Online Social Networks.

[3] O'Neill, P. H. (2019, July 15). How WeChat censors private conversations, automatically in real time.

[4] Xiong, J. K. (2019, July 15). An Analysis of WeChat's Realtime Image Filtering in Chats.

[5] MAITRA, S. (2019, February 24). What Canny Edge Detection algorithm is all about.

[6] Crete-Nishihata, M., Knockel, J., Miller, B., Ng, J. Q., Ruan, L., Tsui, L., and Xiong, R. Remebering Liu Xiaobo: Analyzing censorship of the death of Liu Xiaobo on WeChat and Weibo. Tech. rep., Citizen Lab, University of Toronto, 2017.

[7] Ruan, L., Knockel, J., and CreteNishihata, M. We (Can't) Chat: "709 Crackdown" Discussions Blocked on Weibo and WeChat. Tech. rep., Citizen Lab, University of Toronto, 2017.

[8] Ruan, L., Knockel, J., Ng, J. Q., and CreteNishihata, M. One App, Two Systems: How WeChat uses one censorship policy in China and another internationally. Tech. rep., Citizen Lab, University of Toronto, 2016.

[9] Szegedy, C., Zaremba, W., SUtskever, I., BrUna, J., Erhan, D., Goodfellow, I., and FergUs, R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013).

[10] MathWorks. Convert RGB image or colormap to grayscale.

[11] *Perceptual Hashing*. (2022, October 5). Matt Rickard. https://matt-rickard.com

[12] Renganathan, V., Babu, A.N., & Sarbadhikari, S.N. (2013). A Tutorial on Information Filtering Concepts and Methods for Bio-medical Searching. *Journal of Health and Medical Informatics, 04*.

[13] Lee., "The Porn Breakers," The engineer 291(7610), pp. 30. 2002

[14] Hove, L. J. (2004, April). Extending image retrieval systems with a thesaurus for shapes. In *Norsk Informatikk Konferanse, Stavanger, Tapir Akademisk Forlag*.

[15] Feng, Z., & Tien, D. (2005, July). Enhancement of Semantics in CBIR. In *Third International Conference on Information Technology and Applications (ICITA'05)* (Vol. 1, pp. 744-745). IEEE.

[16] Hsieh, C. J., Liu, W. C., & Li, J. S. (2007, December). An Efficient Packet-level JPEG Forensic Data Collection. In *Future Generation Communication and Networking (FGCN 2007)* (Vol. 2, pp. 108-113). IEEE.

[17] Lin, Y., Tseng, H., & Fuh, C. (2003). Pornography Detection Using Support Vector Machine.

[18] Ibrahim, A. A. (2009). *Detecting and preventing the electronic transmission of illicit images* (Doctoral dissertation).

[19] Bhalerao, D.D., & Parihar, A. (2015). Illicit Image Filtering and Classification Techniques. *International journal of engineering research and technology, 4*.

Cite this article as :

Dr. Manju Bargavi, Sakshi Dhruva, Tenzin Kunsang, S Subham Patra, Tenzin Nyima, "Icensor : Unwanted Image Detection and Censoring", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 2, pp. 75-85, March-April 2023. Available at doi : https://doi.org/10.32628/IJSRSET231027
Journal URL : https://ijsrset.com/IJSRSET231027