Morphological Detection of Common Hate Symbols Using Real World Memes

Sage Fuentes,

Will Custis

Team Number 14

April 25, 2024

Final Project Report for

HCI 575: Computational Perception

Iowa State University

Spring 2024

# ABSTRACT

Online hate and prejudicial imagery via memes have become an undeniable facet in the complex problem of hate, not just online, but offline as well. Despite the benefits that computer vision and image processing can provide to increasing awareness of this problem, there is a surprising lack of research on detection of hate symbols within memes. This study proposes an application of morphological methods to detect three key hate symbols within memes with additional use of OCR for detecting slurs.

## 1 Introduction

Memes have been used as a vector for the proliferation of hate speech in online communities for an indetermined amount of time. Notable websites such as 4chan and X (formally Twitter) have been connected with either the creation or culture of creating prejudice memes [1, 2], though this is not an exhaustive list of all online communities which have contributed to this proliferation.

While we are unsure of the full impact of hateful memes in online spaces, we have seen connections between online hate and real-world events via correlation in increased hate crimes toward individuals on the basis of race, ethnicity, religion, being female, having a non-heterosexual sexual orientation, or identifying as a gender other than your assigned gender at birth [3, 4, 5]. Further, significant events have been facilitated or celebrated by memes. The Christchurch Mosque shootings which involved the death of 51 individuals was carried out by an individual that included references to various memes in his manifesto, specifically memes that were relevant at the time for portions of websites like 8chan and 4chan [6]. Memes have even been an important catalyst in the January 6th Insurrection as it was first discussed abstractly accompanying memes in early 2019 on 4chan’s /k/ or “weapons” board an over the subsequent months [7].

Looking at real world data that coincides with the above rises in internet prejudice, since 2019 hate crimes have increased in the United States by a reported 53.1% (Figure 1) [8]. Looking at monthly frequency, the 2019 highest recorded month for hate crimes (August 2019 at 901 reported offenses) compared to the lowest and highest months in 2023 show that there was a 6.8% increase for the month with the lowest rate (February 2023 at 962) and a 67.8% increase for the month with the highest count (October 2023 at 1,512). This is further complicated by the over 2% decrease in population coverage by the FBI since 2019 and research suggesting that up to 66% of hate crimes go unreported with many counties and cities reporting 0 hate crimes [9].

Figure 1: Hate Crime Frequency Count from 2019 to 2023.

(Data Credit: https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/hate-crime)

These events and statistics set a precedence for a greater need of awareness and detection of hate speech and hate iconography especially within the context of memes that are a part of the complex processes of spreading prejudicial beliefs online. Computer vision and image processing present themselves as useful tools for the organizations and researchers looking to become aware of hateful digital imagery. Despite this, there is a surprisingly small amount of research on computational analysis of hateful imagery and even less on memetic images

The research aims to add to the growing body of work on detecting hate symbols within

images and considers three prominent hate symbols that have been identified by the Anti-Defamation League [10]. The three symbols we have chosen are the Swastika, the Iron Cross, and the SS symbol. These symbols were selected due to their prominence within white-supremacist and Neo-Nazi culture, the relatively context independent nature within the US culture, their identifiability, and their recognizability. Identification of these symbols using purely morphological means is the gap of knowledge that we wish to close.

## 2 Related Work

A notable gap exists around identifying specific hate symbols within memes using morphological means, and to some extent, even around identifying hate symbols using deep learning methods. The current research on this topic lies almost exclusively within deep learning classification and contextual sentiment analysis. An overview of the current state of this line of research has so far focused on either procedurally generated or constructed image datasets, such as the Meta Hateful Memes challenge dataset, or memes collected from various online websites.

### Generated and Constructed Datasets Research

The majority of research in this space is around the Meta Hateful Memes challenge dataset, a dataset of procedurally generated memes for the challenge of contextual sentiment analysis between images and text [11]. Since its creation in 2020, multiple papers have been published on the dataset [a non-exhaustive list: 12, 13, 14]. For image generation, memes were collected from online image sources, filtered to remove specified violent content, and their text was extracted. This text was matched to new images obtained via Getty images and construction completed using an undisclosed tool.

Notably, there is at least one dataset we found where it was created utilizing generative AI tools such as GPT-4V and Stable Diffusion in a dataset created for attempting the Online Safety Prize Challenge (OSPC) [15]. The methods were similar to that of the Meta Hateful Memes images in that the prompts were generated out of collected real world images.

The core difference of procedurally generated or constructed memes and real-world memes constitutes one of the biggest issues in this line of research. Generated memes, or at least the ones based on the procedure for these data sets, cannot accurately replicate the true nature of prejudicial memes from online spaces. They cannot mutate at the rate of online culture in the way that memes do [1].

The Meta dataset omits a long list of violent content and slurs from its content and even ensures that there are appropriate licenses for all images, which is incongruent with the observable culture of memes where there is no consideration for licensing of images in either benign or hateful memes (Figures 2-5). Using generative AI presents additional issues on the basis of collected images for prompt creation, bias and represented knowledge at the researcher and model level, restrictions of the model’s ability to produce images featuring content that violates safety rails, and was even described by the authors of the paper as generating some memes that were “were overly simplistic or failed to recognize specific historical figures in memes.” [15, p. 1895].

Lastly, there are issues with oversampling and gaming the system with generated datasets, specifically noted in Badour and Brown on the Meta Hateful Meme Challenge dataset. Considering accuracy alone for the model representing the top solution of the challenge (at the time of publishing in 2021), the model had an 89% accuracy rate for the Hateful Meme Challenge set which dropped to 73.5% for the Innopolis dataset (cover further in the next section) [16]. While improvements to the Meta Hateful Meme Challenge have seemed to improve from more recent literature, this still leads to serious questions on the useful of generated datasets for complex issues like prejudice and hate.

A skunk in the grass

Description automatically generatedA skunk in the grass

Description automatically generatedA cartoon of a frog

Description automatically generated

(From left to right)

Figure 2: An original meme depicting a skunk with the text, “love the way you smell”, collected for the Meta Hateful Meme challenge dataset [11]

Figure 3: A constructed meme depicting a skunk with the text, “love the way you smell”, from the Meta Hateful Meme challenge dataset [11]

Figure 4: An image of the character Patrick from the cartoon show *Spongebob Squarepants* with the text, “Mother Nature when Father Industrialism walks in”, an example of unlicensed images in memetic content that is benign from a prejudicial perspective.

(Collected from imgur.com, an image aggregate website)

Figure 5: An image of the character Pepe the Frog face “tattooed” with a variety of hate symbols connected to Nazi and Neo-Nazi groups including the three hate symbols of research interest: the Swastika, Iron Cross, and SS symbol.

(Collected from an archived reddit.com post)

### Real-World Dataset Research

In contrast, a smaller part of the field is dedicated to studies involving real world data, memes collected from various websites such as Reddit, 4chan, 8chan, X, or Gab. In addition to the limits this place is on our understanding of real-world data, this also means that we have a limited number of data sets for analysis of hateful and prejudicial memes the following work represents research or compiled repositories on real world data, some of which include their own datasets. The contents of these datasets have been considered for the inclusion in the dataset of our study where permissible and relevant to our study aims.

Development on the MIMIC (Multimodal Islamophobic Meme Identification and Classification) model used real world data to create an identification and classification system specifically for the use case of memes with anti-Muslim prejudice [17]. This study used the Vision-and-Language Transformer (ViLT) model as a basis for their development in the image task. Images were collected from various websites. Despite usefulness in the progression of methods on hateful meme detection, none of the memes from this study were considered relevant for our data set based on lack of relevant hate symbols.

A study by Sun et al. considered the change and mutation overtime of a single meme character, Pepe the Frog, due to his adoption by alt-right and other ideological extremists for prejudicial memes [18]. Multiple tasks were handled by a LLaVA model, a multimodal model with visual analysis capabilities and a chatbot ability to interact. Tasks included using the modal as an interrater for confirming if Pepe was or was not in any given image, generated image descriptions, and OCR. In addition to memes, a multimodal suite of data was assessed, including news articles and blogs that was used to further substantiate the cultural changes with the character that were synthesized with the findings from the model’s analysis of the meme. Much like the MIMIC model, this research needed to consider contextual analysis due to the character of Pepe not being inherently associated with hate speech and requiring an overview of accompanying images and text.

In contrast with the two previous examples that represent the limited recent research on real world images, work by Badour and Brown represents an older attempt to handle hateful memes [16]. Their study is based on the Meta Hateful Memes challenge dataset, but their research also expands to the creation of another dataset of over 23,000 real-world hateful memes, the Innopolis Hateful Memes dataset. This seems to be one of the few datasets of real-world hateful memes contained multiple prejudice types. The methods employed for model creation in the study have been outperformed due to the paper being published in 2021, but the meme dataset provides potential data for the consideration of our research.

Lastly, the only research on computer vision that seems to remotely mention identification of the swastika is by Schinas et al. where visual similarity across several images containing swastikas is displayed (Figure 4 of related paper, p. 662) [19]. No further information is given on the accuracy rate in this application. Additionally, it seems the model used does not detect it as a swastika as it either isn’t referred to in the model’s accompanying text, or in one instance, is identified as a Celtic cross.

### Impact on Our Study

All the research mentioned so far have methods that primarily differ from our approaches due to the extensive use of deep learning models in opposition to our use of pure morphological operations for detection. For symbol identification, image pre-processing in this study will be more intensive due to the need to conform images and find adaptive parameters that are sufficient for multiple types of images. For the dataset, as there does not exist a current dataset specifically for memes containing hate symbol, we have collected images from online sources such as Reddit, Know Your Meme, 4chan, and X and are searching datasets, such as the Innopolis Hateful Memes, dataset for supplemental data.

## 3 Experimental Platform

For this research, OpenCV via Python is being used for its host of image manipulation capabilities. In keeping with the inherent multimodal nature of memes as established in previous literature, text analysis will be considered alongside analysis of the images themselves. OCR will be supplied by an off-the-shelf OCR library such as Pytesseract and additional text detection will be explored via morphological means in cases where text may be difficult to determine via OCR.

An adaptive or algorithmically selected thresholding method for binarization presents itself as a necessity, however early exploration of these methods on a subset of collected data have led to a working understanding that additional steps may be needed before even the binarization step to capture information about memes with vastly different compositions. Figure 6 demonstrates this by showing the meme from Figure 5 binarized via a global threshold and an adaptive threshold. Consider that in the adaptive threshold image that two of the swastikas are no longer connecting due to the quality of the meme. Not pictured here, algorithmic selection via Otsu’s binarization and triangle method were ran on this same image, and despite various efforts, both of these methods provided images with no useful information as the whole of Pepe’s face was made white.

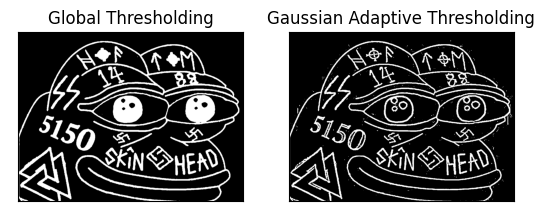


Figure 6: Pepe from Figure 5 with two different threshold approaches applied.

Due to the wide variety of ways in which memes are created, we anticipate needing transformational matrices to reliably detect symbols within the breadth of images. OpenCV’s warpAffine function provides this by allowing us to apply rotations, sheers, and scaling to our detecting structural elements via Affine transformations.



Figure 7: An image of a Swastika with a novel Affine transformation applied. [10]

Two methods for structural elements have been considered thus far, the whole of each hate symbol or symmetrical segments of each hate symbol. Due to the same reasons for the need for the Affine transformation, segments of each hate symbol may need to be considered to deal with warping or partial obfuscation. Segmenting the structural elements may also prove valuable for asymmetric warping as this could be handled by Affine transformations but may prove intractable or computationally inefficient.

## 4 Methodology

### Dataset

The dataset has been curated thus far from posts on sites such as Reddit, X, 4chan, and other image hosting platforms, with some supplemental data from the Innopolis Hateful Meme dataset (though at time of writing, this dataset is still be searched due to it being over 20,000 images in total). Data will be labeled for each hate symbol it has as well as each slur it contains.

As there is not a standardized list of slurs, slurs will be identified from memes themselves and additional slurs for content analysis will be included from observed internet activity to the best of our ability.

### Proposed Pipeline

In the case of both whole symbols and segments of symbols, our proposed pipeline looks the same. Pre-processing for image detection must allow for binarization to effectively work for identification and thus might justify varying categories of meme to better under. Identification of the symbols then must undergo a series of incremental Affine transformations to correctly conform to the symbol present in the image. For OCR processes, documentation from the most well-fitting off-the-shelf library will be followed and detailed included in the full report.

### Evaluation

Our algorithm will be tested with a dataset of 200+ memes. For evaluation purposes, memes will be included in the testing dataset that are not representative of any of the related hate symbols and speech. Success with an individual image will be defined as the algorithm correctly identifying if the image does or does not contain any of the hate symbols or key words we are searching for. When testing the algorithm with the entire dataset, we will strive for 70% accuracy in detecting hate symbols and key terms while keeping the false positive rate below 15%. Based on the results of testing the dataset, we will improve the algorithm depending on the features identified in false positives or negatives. As we can use a smaller sample to inform morphological choices compared to a deep learning model, testing will be performed in batches with iterations being documented as to prevent the morphological equivalent of overtraining.

# REFERENCES

[1] Colley, T., & Moore, M. (2022). The challenges of studying 4chan and the Alt-Right: ‘Come on in the water’s fine.’ *New Media & Society*, 24(1), 5–30. <https://doi.org/10.1177/1461444820948803>

[2] Hickey, D., Fessler, D. M. T., Lerman, K., & Burghardt, K. (2025). X under Musk’s leadership: Substantial hate and no reduction in inauthentic activity. *PLOS ONE*, 20(2), e0313293. <https://doi.org/10.1371/journal.pone.0313293>

[3] Lupu, Y., Sear, R., Velásquez, N., Leahy, R., Restrepo, N. J., Goldberg, B., & Johnson, N. F. (2023). Offline events and online hate. *PLOS ONE*, 18(1), e0278511. <https://doi.org/10.1371/journal.pone.0278511>

[4] Müller, K., & Schwarz, C. (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, 15(3), 270–312. <https://doi.org/10.1257/app.20210211>

[5] Unlu, A., Truong, S., Sawhney, N., Tammi, T., & Kotonen, T. (2025). From prejudice to marginalization: Tracing the forms of online hate speech targeting LGBTQ+ and Muslim communities. *New Media & Society*, 14614448241312900. <https://doi.org/10.1177/14614448241312900>

[6] Thomas, E. (2020). Manifestos, memetic mobilisation and the chan boards in the Christchurch shooting. *Counterterrorism Yearbook* *2020*, 19–22.

[7] Tuters, M. (2021). A prelude to insurrection: How a 4chan refrain anticipated the capitol riot. *Fast Capitalism*, 18(1), 63–71. <https://doi.org/doi:10.32855/fcapital.202101.006>

[8] FBI. *Hate Crime in the United States Incident Analysis*. <https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/hate-crime> (Accessed 3/10/2025)

[9] Jared Shirts. *Improving Hate Crime Reporting Is Necessary for Justice.* [*https://civilrights.org/blog/improving-hate-crime-reporting-is-necessary-for-justice/*](https://civilrights.org/blog/improving-hate-crime-reporting-is-necessary-for-justice/) *(*Accessed 3/12/2025). Note: there is referenced internal research for this statistics, a full report can be found linked in this article, however this was the only article I could find the link to that report in.

[10] ADL. *Hate on DisplayTM Hate Symbols Database.* <https://www.adl.org/resources/hate-symbols/search>. (Accessed 3/05/2025)

[11] Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *Advances in Neural Information Processing Systems*, 33, 2611–2624. <https://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html>

[12] Chen, Y., & Pan, F. (2022). Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17(9), e0274300. https://doi.org/10.1371/journal.pone.0274300

[13] Ivan, S., Ahmed, T., Ahmed, S., & Kabir, Md. H. (2024). A vision-language multimodal framework for detecting hate speech in memes. *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 464–468. https://doi.org/10.1109/CCECE59415.2024.10667078

[14] Velioglu, R., & Rose, J. (2020). Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to Hateful Memes Challenge *arXiv*. <https://doi.org/10.48550/arXiv.2012.12975>

[15] Cao, J., Zhang, Z., Wang, H., Liang, B., Wang, H., & Wong, K.-F. (2024). OSPC: Detecting harmful memes with large language model as a catalyst. *Companion Proceedings of the ACM Web Conference 2024*, 1892–1895. <https://doi.org/10.1145/3589335.3665995>

[16] Badour, J., & Brown, J. A. (2021). Hateful memes classification using machine learning. *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. https://doi.org/10.1109/SSCI50451.2021.9659896

[17] Islam, S. M. J., Mustakim, S. H., Ahmmed, S., Sayeedi, M. F. A., Khandoker, S., Dhrubo, S. T. A., & Hossain, N. (2024). MIMIC: Multimodal Islamophobic meme identification and classification. *arXiv*. <https://doi.org/10.48550/arXiv.2412.00681>

[18] Sun, Y., Pendyala, V., Lian, R., Xin, H., Patel, P., Bucy, E. P., & Shah, D. V. (n.d.). From internet meme to the mainstream: Using computer vision to track “Pepe the Frog” across news platforms. *Visual Communication Quarterly*, 0(ja), 1–25. https://doi.org/10.1080/15551393.2025.2455495

[19] Schinas, M., Galopoulos, P., & Papadopoulos, S. (2023). MAAM: Media asset annotation and management. *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 659–663. https://doi.org/10.1145/3591106.3592232