Motif Mining: Finding and Summarizing Remixed Image Content

William Theisen, Daniel Gonzalez Cedre, Zachariah Carmichael, Daniel Moreira,
Tim Weninger, and Walter Scheirer
Department of Computer Science and Engineering, University of Notre Dame
{wtheisen, dgonza26, zcarmich, dhenriq1, tweninge, walter.scheirer}@nd.edu

Abstract

On the Internet, images are no longer static; they have become dynamic content. Thanks to the availability of smartphones with cameras and easy-to-use editing software, images can be remixed (i.e., redacted, edited, and recombined with other content) on-the-fly, allowing a worldwide audience to repeat the process many times. From digital art to memes, the evolution of images through time is now an important topic of study for digital humanists, social scientists, and media forensics specialists. However, because typical data sets in computer vision are composed of static content, there has been limited development of automated algorithms for analyzing remixed content. In this paper, we propose the idea of Motif Mining: the process of finding and summarizing remixed image content in large collections of unlabeled and unsorted data. For the first time, this idea is formalized and a reference implementation grounded in that formalism is introduced. We conduct experiments on three meme-style data sets, including a newly collected set associated with the Russo-Ukrainian conflict. The proposed motif mining approach is able to identify related remixed content that, when compared to similar approaches, more closely aligns with the preferences and expectations of human observers.

1. Introduction

As the number of images posted online has grown, it has become increasingly intractable for humans to manually discover trends in online information. Although some computer vision algorithms have been proposed for this problem [31, 36], the financial, labor, and time costs associated with labelling ground truth present a tremendous hurdle. This is particularly evident when analyzing social trends, which often move so quickly that well-labelled data becomes obsolete by the time it is prepared. Moreover, the prevalence of remixed image content like memes—images often edited to remove information or incorporate other content—has raised questions about how associate related images. In this paper, we describe the process of automatically discovering trends in a large collection of remixed images, known as *Motif Mining* (Fig. 1). Several different communities are interested in this concept, including digital humanists studying new participatory art movements [25], computational social scientists studying the role of visual communication in conflicts [32], and media forensics specialists attempting to detect disinformation [16].

There are several challenges that must be overcome to achieve robust and accurate motif mining. To date, this concept has been applied informally in the literature [35, 3, 6, 29], leaving questions about optimization strategies that can be applied to the problem and the structure of the output. With respect to a viable algorithm that expresses image similarity via a graph, no image feature exists that works with both globally similar images and images that are similar only in small local regions—what is commonly observed in remixed content. Additionally, there has yet to be a large study of how different combinations of image features and graph building algorithms affect the human perception of mined motifs. As the purpose of motif mining is to aid human observers, this is an important question to answer.

This paper makes the following contributions:

1. A formal description of the problem of motif mining, providing a roadmap on how to more-easily discover salient trends in large, unsorted data sets, accompanied by a solution to this problem in the form of an end-to-end processing pipeline.[1]
2. A new method for generating motif graphs leveraging vector retrieval systems that shows increased accuracy and speed over prior work.
3. A new image feature strategy for this problem combining both local and global image features to increase the context available during graph generation.
4. A new data set of over half of a million posts containing remixed and static images collected from Telegram over the past six years, including the beginning of the 2022 invasion of Ukraine by the Russian Federation.
5. An empirical study of the proposed pipeline, including

[1] This system will be open-sourced pending publication of this work.

Figure 1. Given a large, unsorted, and unlabelled data set from social media, the motif mining strategy arranges related content—especially remixed content—into interpretable graphs in an unsupervised manner. Above are two motifs mined from a new data set of images related to the Russo-Ukrainian conflict collected from the Telegram messaging platform [28].

comparisons to related approaches on three data sets containing remixed and static image content.

2. Related Work

Within computer vision, motif mining is most closely related to content-based image retrieval (CBIR) [7]. CBIR takes as input (i) a query image of interest and (ii) a gallery of potentially related images and aims to retrieve the images from the gallery that are similar to the query, sorting them into a ranked list from the most to the least similar according to well-defined similarity criteria. Depending upon the CBIR system user's intent, the similarity criteria may range from retrieving images that are semantically similar (*e.g.*, images that depict the same type of objects as the query) to retrieving images that are near-duplicates (*e.g.*, images that are minor variations of the query, thus sharing many pixels that come from the same imaging pipeline).

Features used for matching images in CBIR can be local or global and handcrafted or learned. Popular examples of handcrafted local features useful for retrieving semantically similar images include SIFT [14] and SURF [2] while more recent learned local feature approaches include LIFT [34], DELF [18], and LISRD [21]. Handcrafted global features that are useful for finding near-duplicate images, in turn, involve concatenating patch-wise features such as LBP [19] and PHASH [15]. More recently, the use of intermediary convolutional layers of neural networks as global image descriptors has become popular, including features from VGG [26], ResNet [10], and MobileNet [11] (hereafter MOBILE). These features work for the retrieval of semantically similar images because they leverage the learned object classification ability of their respective networks.

CBIR solutions index low-level features to reduce storage costs through feature compression and to speed up the retrieval of feature-wise k-nearest neighbors. Standard feature indexing is based retrieving approximate nearest neighbor (ANN) features for each of the query's features, supported by optimized product quantization (OPQ [8]) of all the features. FAISS [12] is a popular library that implements different indexing strategies, including OPQ.

Finally, at the highest level, once a set of features from the gallery images is retrieved for all of the query's features, a voting scheme leveraging an Inverted Vector File (IVF) is used to find the gallery images that are the most similar to the query. Using the voting count, gallery images can be sorted from most to least similar, providing us with the desired output: a ranked list of images similar to the query.

In contrast to CBIR, motif mining takes as input a large data set of images of interest only; there is no targeted query to take as a reference. Moreover, rather than returning a ranked list of similar images, the purpose of motif mining is to find different motifs, *i.e.*, graphs of images whose similarity is not known at execution time. Since these similarities are sometimes semantic (in the case of conceptually similar images) and sometimes based on pixel values (in the case of images sharing templates, such as memes containing stock character macros [25]) or even both (such as the motif on the left-hand side of Fig. 1), they are to be discovered directly from the data.

Despite their differences in both input and purpose, motif mining uses as its base the best indexing strategies from CBIR pipelines. With respect to motif mining's output, an interpretable graph that represents the relationships between images is constructed, similar in spirit to the clustering methods proposed for the specific case of image-based memes by Zannettou et al. [35], Beskow et al. [3], Dubey et al. [6], and Theisen et al. [29]. Our approach, however, is not constrained to just remixed content like memes—it can identify visually similar static images in an unsupervised manner as well.

The work of Theisen et al. [29] is the most related to this paper. However, there are three key differences. (1) Theisen et al. explicitly noted that feature-level fusion was a possibility, but did not incorporate it. In this paper we propose

a new method for this by appending a small piece of global context to each local feature for an image. (2) Theisen et al. made no guarantee that the produced graph contained no isolated vertices, and the exclusive use of Spectral Clustering [17] required that a number of clusters be specified beforehand. (3) Compared to the work of Theisen et al., our approach is an order of magnitude faster in processing time, making it more suitable for analyzing large galleries. This was achieved by moving both the feature extraction and index operations onto a GPU.

3. Formalization of Motif Mining

Traditionally, computer vision problems have been framed as an optimization over some metric calculated in reference to ground truth data for a task. Although this provides high-quality baselines for comparison, there is often little effort expended on demonstrating that higher metric scores actually result in more useful output for human observers for tasks like image retrieval. As an alternative, the methods and procedures of visual psychophysics from psychology have been recommended as a way to use human behavioral responses to evaluate algorithms [23, 24]. Taking insight from that work, we formalize motif mining.

The Motif Mining Problem. The purpose of motif mining is to allow human observers to quickly gain insights about visual trends in a large collection of unsorted and unlabeled data. A common method for finding multiple trends in a given data set is via graph building. However, as the purpose of motif mining is to aid people, the graphs must be optimized around some human feedback mechanism. We structure our experiments around this idea.

For an example of a useful graph, the right-hand side of Fig. 1 shows a number of different airplanes, several of which are fighter jets. This example is drawn from the current Russo-Ukrainian conflict. An increase in the number of militaristic images being posted online might prefigure an event in a conflict [32] and could also potentially leak useful and/or damaging intelligence to third parties. The Ukrainian government recently addressed this concern, specifically, with "Ukraine's defense minister, Oleksii Reznikov, [...] calling on viewers to share images of Russia's assault" and "a local Telegram channel urged its 400,000 subscribers to 'carefully film' and share video of passing Russian troops so Ukrainian fighters could hunt them down" [9]. These examples were taken from 851 motifs mined from a subset of 16,433 images from the Ukrainian data set collected from Telegram [28]. The ideal number of graphs and the distribution of the images across them is best formalized as an optimization problem with task accuracy being derived from human feedback.

Optimization for Human Observers. Given a large, unlabelled data set of images, our goal is to automatically discover trends in it by classifying those images that can be thought of as being "conceptually similar" or "derived from the same picture" in some intuitive sense. Because our task is both inherently subjective and difficult to formally specify, and because the quantity of data far exceeds any human annotators' ability to manually label, we develop an unsupervised system for grouping these images together and verify them *a posteriori* with human aid.

Our formal framework specifies a data set of images as a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w : \mathcal{E} \to \mathbb{R}_+)$, where $\mathcal{V}$ is the vertex set for the graph $\mathcal{G}$ and $\mathcal{E}$ is its edge set. Here, $w : \mathcal{E} \to \mathbb{R}_+$ denotes a function that assigns positive, real-valued weights to each of the edges of $\mathcal{G}$. The vertices in this graph represent images from the data set, and the weighted edges capture the strength of the similarity between two adjacent images. Within this framework, the task becomes computing an unsupervised clustering of $\mathcal{V}$ that disagrees as little as possible with what human observers expect. We test this using the *Imposter-Host Task* [30]. This means finding some partition $\mathcal{C}$ of the vertex set $\mathcal{V}$ such that, for a given pair of distinct clusters $c, \tilde{c} \in \mathcal{C}$, if a human were presented with k images from c and one imposter image from $\tilde{c}$, the human would be able to pick the imposter (*i.e.*, the odd image out). We define this formally as follows:

$$\min_{\mathcal{C} \in \mathfrak{P}(\mathcal{V})} \left(\sum_{c \neq \tilde{c} \in \mathcal{C}} \; \sum_{v_1, \ldots v_k \in c} \; \sum_{\tilde{v} \in \tilde{c}} \left(1 - H(v_1, \ldots v_k, \tilde{v})\right) \gamma(c, \tilde{c}) \right), \tag{1}$$

where (i) $k \in \mathbb{N}_+$, (ii) $\mathcal{G}$ is our weighted graph, (iii) $\mathfrak{P}(\mathcal{V})$ is the set of partitions of $\mathcal{V}$, (iv) $\mathcal{C}$ is a set of vertex clusters, (v) $v_1, \ldots v_k$ all belong to the same cluster c, (vi) $\tilde{v}$ belongs to a different cluster $\tilde{c}$, (vii) $H : \mathcal{V}^{k+1} \to \{0, 1\}$ returns 1 *iff* $\tilde{v}$ is correctly identified by a human, and (viii) $\gamma(\cdot, \cdot)$ is a normalizing factor (see Supp. Mat. Sec. I).

In order to specify this graph $\mathcal{G}$, a mapping process $\mathfrak{M}$, with some corresponding parameters, is needed to map the image corpus onto a weighted graph. With this in mind, the problem can be further thought of as an optimization task like Eq. (1) for each weighted graph realizing the given data set. Thus, Eq. 1 will be minimized for a given clustering $\mathcal{C}$ of $\mathcal{G}$ only when human observers agree with the quality of the clustering. Simpley enumerating all possible clusterings is obviously computationally infeasible; instead, it would be more principled to parameterize a clustering algorithm A and perform this optimization task over A's set of parameters. However, this would require an enormous number of human observers to check each clustering.

In this paper, we employ a variety of effective graph clustering algorithms and check their performance directly against the human observers for a few different realizations of the graph produced by $\mathfrak{M}$. Every time $\mathfrak{M}$ produces a weighted graph, we apply one of the clustering algorithms A_i to the graph and evaluate the quality of those clusters. This heuristic approach is a step in the direction of finding

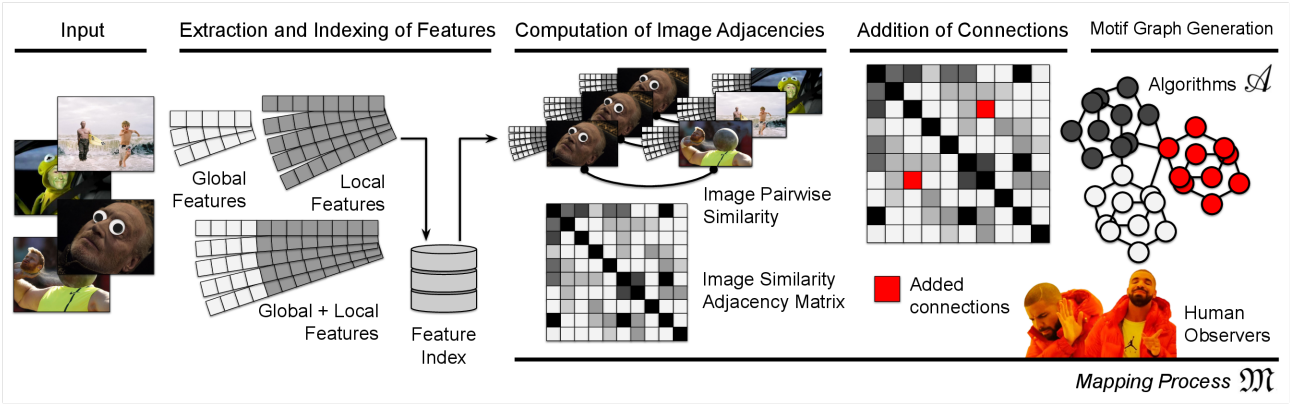

Figure 2. Motif mining pipeline. The process starts with a large set of images of interest as input and ends with the mined motifs, *i.e.*, graphs of the images that summarize remixed content. Four steps constitute the pipeline: extraction and indexing of low-level image features, computation of image adjacencies according to their pairwise similarities, computation of additional similarity connections, and building graphs of images that represent the motifs. Human observers provide feedback on the quality of the motifs for evaluation.

a true minimum to the optimization in Eq. 1.

4. Implementation of Motif Mining

Fig. 2 summarizes our pipeline for motif mining. Each of the steps depicted within it is detailed below.

Extraction and Indexing of Features. The first step towards motif mining is to determine the kind of features that will be used to generate the vectors in the index. This decision is informed primarily by the types of motifs an observer wishes to find in the data set. Global and local feature extraction methods will produce intuitively different results in the types of images returned by a given query and therefore the types of motifs mined. PHASH [13] features, for instance, will return only duplicate or near-duplicate images. MOBILE [11] and VGG [26] features will return images that are similar globally in a semantic sense; this often manifests as something akin to images that all contain airplanes, though those airplanes may be different shapes, sizes, positions, and styles. SURF [2] features lead to connections that may be visually diverse and share only some very small local information, such as a logo on a flag.

In this work, we show that it is the combination of these two types of features that creates compelling and robust connections. Quite frequently, SURF features will not return near duplicate images in the top of their query results as there are smaller, more subtle matches somewhere in the image. However, global visual similarity is very apparent and useful to human observers. If, for example, military groups begin to post edited pictures of tanks or inflammatory extremist symbols, human observers will want to quickly identify these trends. To capture both of these cases, each image has a single global feature extracted in addition to its SURF features. The full set of global features is then subjected to Principal Component Analysis (PCA) [22] and each vector is reduced to 16 dimensions to match the length of the smallest of the descriptors (namely, PHASH). A further discussion about this process is provided in Sec. E.1 of the Supp. Mat. This global descriptor is then appended to each of the respective images' SURF features. Therefore, global context for any given image is incorporated into all of its individual local feature vectors. The effect of adding this global context can be seen in both the airplane motif on the far right of Fig. 1 and the flag motif in the third row and third column of Fig. 3. The local features come to the fore when the motif is a small object in an image, such as the Indonesian ballot boxes with "KPU" written on them as seen in second row and third column of Fig. 3.

To index the features, the proposed pipeline is implemented using the FAISS [12] library as a foundation. Available within FAISS, OPQ [8] allows for efficient mass-vector indexing and retrieval (for time-complexity considerations, please see [8]). Similar to the work previously done by Theisen et al. [29], an IVF index is made with 256 centroids (an exploration of how the number of centroids affects the index and resulting graphs can be found in Sec. E.1 of the Supp. Mat.). This index, once built, provides the function used to generate the graph that is then mined for motifs. Readers familiar with the workings of OPQ may at this point wonder why the centroid clusters that are inherent to OPQ, and already generated by FAISS, are not just used for the end product without the extra hassle of producing another graph on top of the index. This is discussed in Sec. 5. Building this index allows us to quickly construct an approximate affinity matrix, and a subsequent graph, by leveraging the efficiency of FAISS.

Computation of Image Adjacencies. We need to perform pairwise image comparisons to compute image similarities prior to establishing the motif clusters. To compute similarities between images leveraging the index built in the previous step, we select a set of images used as starting points to query features from the index. Given that IVF

indices are built at the local feature-level, the indexed features have a many-to-one relationship with their respective source images. As a consequence, retrieved features need to be "mapped" back to the image from which they were extracted, since we are interested in image-level similarity. Considering the querying of the features of a selected starting-point image, each result $r \in \mathcal{R}$ is a tuple (f, i, d), where f is a feature corresponding to an image i, and d is the distance between f and the queried feature as computed by the index. If we focus on the subset $\mathcal{R}_i$ of the retrieved features that belong to image i, we can compute the similarity s_i between that image and the selected starting-point image as follows:

$$s_i = \sum_{(f,i,d) \in \mathcal{R}_i} 1 - \tanh(d). \tag{2}$$

We elect to use the nonlinear operator $\tanh(\cdot)$ as it is nicely bounded within the interval $[0, 1)$ for all non-negative d. Intuitively, the nonlinear weighting rewards smaller distances and penalizes distances more harshly as they become larger.

Each feature vector difference, after applying Eq. 2, is then added to an image-level score for the image that generated the retrieved feature vector. We can then take the N highest-scoring images, with N being a free parameter (set to 50 in our experiments below), and return a ranked list of the most similar images to any given query image. We call this the "voting" step because images receive "votes" in the form of individual vector distances.

The similarity computation is a loop only for the local features (including the ones combined with global features). For the use of purely global features, since their relationship is one-to-one with the source image, we simply take the single distance value d returned by the index and compute the similarity score $1 - \tanh(d)$ directly instead of using Eq. 2.

To realize a graph out of image similarity computations, we create an $N \times N$ adjacency matrix where each of the N images in the data set determines one row and column. We then fill entry (i, j) of this matrix with the computed similarity score between those corresponding images. The entries in this matrix will then correspond to weighted edges in our final graph between the vertices representing those images in the data set.

Several strategies have been explored for selecting the starting-point images and filling in the scores in this matrix. Prior work [29] has simply taken a smaller subset of the images in the set and hoped that the resulting connections are diverse enough to form a representative graph. In this work, we instead continue selecting a random subset of isolated images in the graph (*i.e.*, images whose columns and rows within the adjacency matrix sum up to zero), until all the images are visited. This method is cheaper than querying all N images while still eliminating any isolated vertices, unlike prior work. Note that this does not ensure one singular connected component, though some features (*e.g.*, SURF) do still lead to fully-connected graphs on smaller data sets.

Addition of Image Connections. Querying the index to generate the graph ensures that there are no isolated vertices but does not ensure that the graph is fully connected. Many clustering algorithms, such as Spectral Clustering [17], work best on graphs with one connected component. Therefore, after the query step, three heuristic-based methods were tested for connecting the graph by addition of extra edges. However, these methods were costly in terms of compute-time and delivered only minor increases in accuracy, seemingly at random, when compared to simply clustering on the unconnected graph. For this reason we believe that fully connecting the graph is not worth the time required to do so. Details of the three different connection methods are available in Sec. A of the Supp. Mat.

Clustering of Images. Three clustering techniques, well-suited to finding communities in weighted graphs, were tested. Louvain clustering, Markov clustering, and Spectral clustering. Spectral clustering was chosen so that results can be compared to the prior literature [29].

The *Louvain* [4] method for community detection maximizes the modularity of the graph—a measure comparing the density within and across clusters—using a two-stage iterative optimization. *Markov* clustering [5], on the other hand, is a random-walk based clustering algorithm that computes transition probabilities between the vertices of a weighted graph by modeling random walks over the graph as Markov chains. Finally, *Spectral* clustering refers to a very popular approach to clustering data according to the eigenvalues of a Laplacian defined over the data. In the context of clustering for graphs, the Spectral approach involves applying k-means clustering to the vertices of the graph using the k largest eigenvalues of the graph's Laplacian matrix $L = D - A$ as features, where A is the graph's (weighted) adjacency matrix, and D is its diagonal.

These clustering algorithms provide an unsupervised way of exposing underlying trends in data—the way in which remixed and static images would naturally be agglomerated by a human. However, which of these, if any, most closely matches human intuition can be revealed only by incorporating human feedback into the evaluation.

5. Experiments and Results

In total, 252 different configurations of the Motif Mining pipeline described in Sec. 4 were tested in order to identify the most effective ones. Fig. 4 provides a summary of all of these results, while Sec. C of the Supp. Mat. presents them individually in a more detailed tabular form. Of primary interest is each particular combination's accuracy on the Imposter-Host test, which serves as a proxy for whether or not the produced motifs are visually recognizable to human observers. To describe the configurations, we use the format

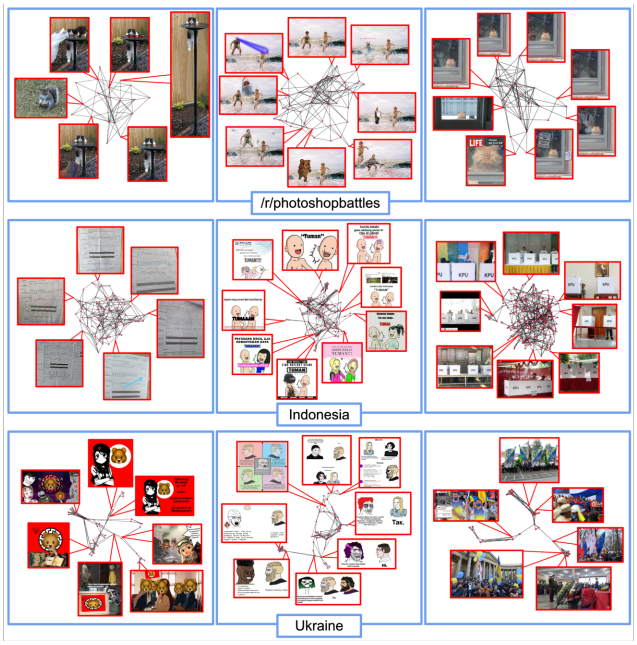

Figure 3. Nine motifs discovered from the Reddit Photoshop Battles data set [16], 2019 Indonesian National Election data set [29], and a newly collected data set associated with the Russo-Ukrainian conflict. A variety of global motifs, local motifs, and remixed content can be seen across the nine examples. From cats to alleged voting fraud, the new pipeline can discover a diverse array of motifs in any data set. Zoom in on PDF to see detail.

`feature_type-connect_type-cluster_type`, where `feature_type` is one of: PHASH, MOBILE, VGG, SURF, SURF_PHASH, SURF_MOBILE, SURF_VGG. `Connect_type` is one of: avg (average), best, er (Erdős-Rényi), or reg (unconnected). Finally, `cluster_type` is one of Louvain, Markov, or Spectral.

Data Sets. The first benchmark data set used for experiments was the *Reddit Photoshop Battles* [16] data set, an image remix-specific set for content-based image retrieval and image clustering. It consists of 10,586 images taken from threads on Reddit's r/photoshopbattles subreddit. It proves to be a particularly challenging collection due to the diversity of the submitted images as seen in Fig. 3. This data set provides the "purest" collection of remixed images, as the purpose of the subreddit is to take a piece of a "donor" image and insert it into a variety of different sub-images. The second set consisted of 44,612 images (including memes) related to the 2019 Indonesian Presidential election (shortened to "Indonesia"), as used in [29], to allow for an additional point of comparison to previous work.

A new data set of interest to the human rights community was also collected (referred to as "Ukraine"). Scraped from Telegram starting in the year 2016 and continuing through the beginning of March 2022, it focuses on content related to the Russo-Ukrainian conflict. Containing controversial images relating to nationalism, fascism, xenophobia, racism, homophobia, and the growth of militia and paramilitary groups, it provides an unprecedented look into the growth of online tensions surrounding the conflict. Comprised of 665,725 images and 721,441 posts, it is relevant for evaluating tools aimed at aiding human rights activists in the fight against online hate and disinformation. Out of this, a subset of 16,433 images was selected as a test set for our experiments.

With the help of experts within Ukraine, a list of Telegram users was compiled. The channels associated with these users were then scraped, forming the data set. A list of these users may be found in Sec. G of the Supp. Mat. Both the data set and the scraping tools will be released alongside the publication of this work. Each post consists of a JSON file containing, in addition to the post title and text, relevant meta-data such as the post date, view count at the time of scraping, image links associated with the post (Telegram allows more than one image per post), and the raw image files.

Assessing Cluster Relatedness with an Imposter-Host Test. As the goal of the pipeline is to create clusters for humans to review, testing the "accuracy" of the clusters requires human input. To this end, we use a version of the *Imposter-Host* test [30] outlined in Theisen et al. [29], a standard way of evaluating classification tasks in a human-centric way. This test consisted of asking 50 Amazon Mechanical Turk [1] workers each to find which image out of a set of five was the most different 25 times separately (with 5 of those 25 being control questions). Four of the images shown were from a single "host" cluster ($k = 4$ in Eq. 1), and the fifth image—that the Turk workers are tasked with identifying—was taken from a randomly selected "imposter" cluster. Intuitively, the more related the images in any given cluster are, the easier it should be for the Turk workers to find the imposter image. Since selections are made from a set of five images, the baseline accuracy (computed by randomly picking one of the five images) is approximately 20% across all trials. This experiment was run for each of the 84 different possible combinations of feature types, connection types, and clustering methods and was done for all 3 data sets. Due to our belief that the distribution of images amongst the clusters matters, and consistent with prior work [29], the accuracy scores have been normalized against their cluster sizes on a per-cluster basis, before aggregating for a per-method accuracy score.

Reddit, the smallest data set, seems to imply that a global feature yields the best results in terms of observer accuracy per Fig. 4. However as explained in Sec. 4.1, the global features are constrained by the number of centroids that the index is initialized with. For a smaller data set like Reddit, containing only 10,588 images, 256 clusters is enough to achieve high accuracy scores (this intuition is based on there being 186 Reddit threads comprising the data set, which, if

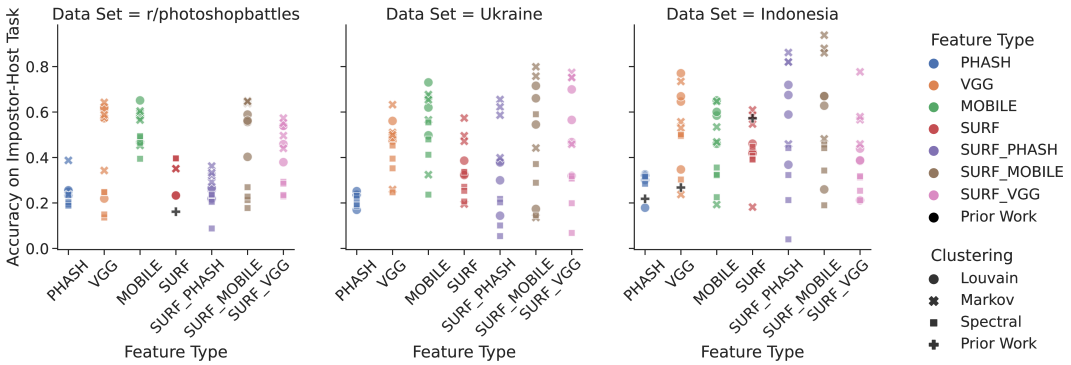

Figure 4. The accuracy scores of the Imposter-Host test across the three data sets. Each of the three clustering methods is noted with a different shape. The results noted as prior work are from Theisen et al. [29] using Spectral clustering on an unconnected graph in different feature configurations. For further details on the connection algorithm used in our work, see Sec. C of the Supp. Mat.

taken as a proxy for classes, implies that 186 mined motifs would be the perfect answer). However, as the data sets grow larger, the available number of clusters stays at 256. This is in contrast to the combined global-local features, which continue to grow in the number of clusters (though not strictly proportionally). In this work, increasing accuracy scores are seen amongst the global-local features as the number of images increases. Should the number of images continue to increase, it seems likely that the accuracy of the global features would begin to decrease as more increasingly visually diverse images will have to be binned in a maximum of 256 clusters, while the global-local features allow for more clusters as the data grows.

With respect to the Imposter-Host task accuracy scores, we achieved state-of-the-art performance. As seen in Table 1, prior literature on the Reddit data set achieved only a 16.15% accuracy, worse than random chance, claiming that "due to the visual complexity of the data set, they [the Turk Workers] were able to find connections that weren't intended to link images" [29]. To show that a more principled method of graph generation alone can greatly improve results, we ran an ablation study, the results of which can be seen in columns 1 & 2 of Table 1. In this comparison, the only difference between the two methods was how the graph was generated. For 5 out of the 6 comparisons, the new graph generation method greatly increased the accuracy scores. The accuracy can be increased further, even when using the same features, by using a different clustering algorithm (*i.e.*, Louvain or Markov clustering) better suited to the problem at hand. In addition to these two improvements to the accuracy when compared to prior work, using the newly-proposed combined feature approach from the present work yet again leads to increases in accuracy. The `surf_mobile` features result in the highest accuracy scores seen in this study. By combining all three of these improvements, accuracy scores can be increased by upwards of 50 percentage points when compared to the previous state-of-the-art (48.96).

On the new 'Ukraine' data set, the accuracy scores are similarly high (Fig. 4, middle plot). A top score of 79.91% is achieved by the `surf_mobile-er-markov` method. In this case, combined features show high accuracy even for larger data sets.

There is an important caveat to this result. Much as Theisen et al. [29] found that Spectral clustering produced a singular "mega-cluster" that contained the vast majority of the images and thus made the method undesirable, the Markov clustering method had a similar issue that inadvertently skewed the results towards the higher side: instead of placing all of the outlier images into a single massive cluster, it placed outliers into their own individual clusters. With the Imposter-Host test requiring at least 4 images in a host cluster, these individual-image outlier clusters were ignored. This left a number of clusters containing only near duplicates, which would thus improve the Imposter-Host accuracy scores. Due to this quirk, unless one is only interested in small near-duplicate clustering, we would recommend using the Louvain clustering method in a real-world implementation of motif mining as it results in a much more even and practical distribution of images and cluster sizes while maintaining state-of-the-art accuracy on the Imposter-Host task (a plot showing this distribution may be seen in Sec. D.2 of the Supp. Mat.).

Underlying Graph and Cluster Structures. Although the new graph creation method ensures that there are no isolated vertices in the graph, it is not guaranteed to produce a single connected component. Instead, the graph contains several interesting patterns emerging from the type of feature used in the creation of the index, and from choices relating to the initialization of the index.

To explore this further, an experiment was run in which each index was recreated on the Reddit data set with a different number of centroids (128, 256, 512, 1024). The global feature types always resulted in graphs with a number of components equal to the number of centroids the index was initialized with. The global features resulting in

	[29] (Top)	Ours (Top - Spectral)	Ours (Top)
Reddit - PHASH	N/A	23.53%	**38.73%**
Reddit - VGG	N/A	24.79%	**64.25%**
Reddit - SURF	16.15%	**39.62%**	**39.62%**
Indonesia - PHASH	21.83%	31.81%	**32.53%**
Indonesia - VGG	26.79%	50.07%	**77.05%**
Indonesia - SURF	57.25%	44.66%	**60.94%**

Table 1. The top accuracy scores on the Imposter-Host task compared against Theisen et al. [29]. Shown are the top scores from the proposed pipeline using the Spectral clustering method to allow for a fair comparison to how the clustering takes place in previous work, followed by the top score for any other combination of features and clustering approach. This shows that building the graph in a more principled manner while still using Spectral clustering can improve the accuracy compared to prior work, but switching from Spectral clustering to a different clustering method can almost always improve scores even further (Reddit-SURF combination aside).

connected components equal to the number of centroids implies that all the querying step of the pipeline is doing is exposing the latent centroid-space that FAISS has already prepared and therefore could be done-away-with entirely (see Sec. E.1 in the Supp. Mat. for details).

Although SURF features by themselves resulted in a highly connected graph, and global features resulted in simply mirroring the underlying cluster space FAISS had already computed, the combined global + local features resulted in a higher number of components than the number of centroids, implying that further sub-structures of similar images were discovered within the centroids. Those combined features producing more components percolate down the pipeline to the clustering step, where the clusters using these features give a better image-cluster spread.

Qualitative Results. During the processing of the Ukrainian data, an initial test of the pipeline was run with a `surf_mobile-reg-louvain` configuration, which we recommend as the best option when running in-the-wild due to its combination of speed, accuracy, and image distribution. The results were interesting enough that we proceeded to run it on all three data sets, producing the results seen in Fig. 1 and Fig. 3. In Fig. 3, the leftmost cluster in the top row shows five images of a cat drinking out of a birdbath and a spurious match on a squirrel with a briefcase. The middle motif is of a man chasing a child out of the ocean. Here, we see much stronger examples of remixing, with laser beams, sharks, and a bear all being added. The final motif is of a cat sitting at a door. This grouping highlights the usefulness of the local matching as one of the images has a stark global contrast from all of the others, but the local features allow the matching on the shared cat's head.

The second row illustrates three motifs from the Indonesian data set. Again, we can see a strong cluster of remixed content in the middle. On the left can be seen images of voting tallies, which the Prabowo campaign used as alleged evidence of fraud in a failed attempt to contest the 2019 election [20]. The third cluster again demonstrates why local features are extremely useful in motif mining: many different images of Indonesians at ballot boxes were present in this motif but only shared the locally similar 'KPU' logo on the boxes.

Finally the bottom row of Fig. 3 shows results from the Ukrainian data set. A cartoon bear head used as a logo by one of the extremist meme channels is found in the left-most panel. Note that this imagery is often superimposed over a Sonnenrad, which is a co-opted Nazi rune [27]. In the middle are several variations of a Ukrainian version of the Yes-Chad meme [33]. On the right is a cluster of Ukrainian coloured flags. Of particular interest is the photoshopped image top-center, showing school children bearing a number of flags. See Supp. Mat. Sec. F for additional examples.

6. Conclusions

The newly proposed pipeline, combined with a novel graph generation technique and combination of image features, achieves state-of-the-art results on the motif mining problem. With increases reaching nearly 50 percentage points improvement over previous work, it demonstrates a path forward for aiding human responses to emerging trends in online social media. In addition, a new data set has been collected from Telegram to allow further benchmarking in this space. Its timely release will allow researchers to gain unparalleled views into the increasing tensions online between Ukrainian and Russian actors, mirroring the growing tensions happening on the ground.

Limitations and Future Work. The proposed work still has several areas for improvement. First, calculating the accuracy of visual motif mining pipelines is costly and time consuming, requiring large amounts of human input. A computational model of the human-feedback metric, possibly making use of modeling work in visual psychophysics from psychology, could decrease testing time and monetary costs. Second, the current pipeline makes no use of non-visual data collected alongside the images (*e.g.*, time-of-post and text collected with the image). Multi-modal approaches to motif mining will surely yield more accurate results. We believe that using all available context is of the utmost importance for future studies in this area.

References

[1] Amazon Mechanical Turk. https://www.mturk.com/.

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[3] D. Beskow, S. Kumar, and K. M. Carley. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing and Management*, 57(2), 2020.

[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[5] Stijn Dongen. Graph clustering by flow simulation. *PhD thesis, Center for Math and Computer Science (CWI)*, 05 2000.

[6] A. Dubey, E. Moro, M. Cebrian, and I. Rahwan. Memesequencer: Sparse matching for embedding image macros. In *In Proceedings of the International World Wide Web Conference*, 2018.

[7] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5), 2021.

[8] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2946–2953, 2013.

[9] Drew Harwell and Rachel Lerman. How ukrainians have used social media to humiliate the russians and rally the world. https://www.washingtonpost.com/technology/2022/03/01/social-media-ukraine-russia/, 2022.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam. Searching for mobilenetv3. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[13] E. Klinger and D. Starkweather. pHash: The open source perceptual hash library. https://www.phash.org, 2013.

[14] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[15] Vishal Monga and Brian L Evans. Perceptual image hashing via feature points: performance evaluation and tradeoffs. *IEEE Transactions on Image Processing*, 15(11):3452–3465, 2006.

[16] Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin Bowyer, Patrick Flynn, Anderson Rocha, and Walter Scheirer. Image provenance analysis at scale. *IEEE Transactions on Image Processing*, 27:6109–6123, 08 2018.

[17] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

[18] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *IEEE International Conference on Computer Vision*, pages 3456–3465, 2017.

[19] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[20] Richard C. Paddock. Indonesia court rejects presidential candidate's voting fraud claims. https://www.nytimes.com/2019/06/27/world/asia/indonesia-widodo-prabowo-election-fraud.html, 2019.

[21] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *European Conference on Computer Vision*, pages 707–724, 2020.

[22] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[23] Brandon RichardWebster, Samuel Anthony, and Walter Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2280–2286, 2018.

[24] Brandon RichardWebster, So Yon Kwon, Christopher Clarizio, Samuel E. Anthony, and Walter J. Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[25] Limor Shifman. *Memes in Digital Culture*. The MIT Press, 2013.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[27] Sonnenrad. https://www.adl.org/education/references/hate-symbols/sonnenrad.

[28] Telegram FZ LLC and Telegram Messenger Inc. Telegram. https://telegram.org/.

[29] W. Theisen, J. Brogan, P. B. Thomas, D. Moreira, P. Phoa, T. Weninger, and W. Scheirer. Automatic discovery of political meme genres with diverse appearances. *AAAI Conference on Web and Social Media*, 15:714–726, 2021.

[30] Tim Weninger, Yonatan Bisk, and Jiawei Han. Document-topic hierarchies from document graphs. In *ACM International Conference on Information and Knowledge Management*, pages 635–644, 2012.

[31] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4066–4075, 2019.

[32] M. Yankoski, W. Theisen, E. Verdeja, and W. J. Scheirer. Artificial intelligence for peace: An early warning system for mass violence. *Towards an International Political Economy of Artificial Intelligence*, pages 147–175, 2021.

[33] Yes Chad. https://knowyourmeme.com/memes/yes-chad.

[34] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483, 2016.

[35] S. Zanneettou, T. Caulfield, J. Blackburn, E. D. Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *ACM Internet Measurement Conference*, 2018.

[36] Junjie Zhao, Donghuan Lu, Kai Ma, Yu Zhang, and Yefeng Zheng. Deep image clustering with category-style representation. In *European Conference on Computer Vision*, pages 54–70, 2020.