Music and Video Alignment in the Ukrainian and Russian Social Media Space

Adam Colton

University of Utah

Abstract

Telegram is a popular platform for sharing news and video media related to the war in Ukraine. Video data was collected from 28 Telegram war-related channels and clustered by similarity to create equivalence frames. A deeper analysis of patterns of music editing inside of clusters of related videos revealed a gradient of channel affinities.

Contents

1 Introduction **1**

2 Background **2**
2.1 The War in Ukraine . . . 2
2.2 Telegram . . . 2

3 Related Work **4**
3.1 Music Emotion Recognition . . . 4
3.2 Research in the Telegram information sphere . . . 4
3.3 Video Similarity . . . 4

4 Ethical Data Collection: Privacy and Safety **4**

5 Methods **6**
5.1 Channel Selection . . . 6
5.2 Telegram Data Collection . . . 7
5.3 Music Identification Pipeline . . . 7
5.4 Video Similarity Scoring . . . 7
5.5 All-Pairs Video Incident Similarity Computed over a Web-Scale Dataset . . . 7
5.6 Future Performance Enhancements . . . 8
5.7 Clusters of Similar Videos . . . 9
5.8 Video Clustering Failure Points . . . 9
5.9 Song Matching Failure Points . . . 9
5.10 Manual Labelling . . . 9

6 Results **13**
6.1 Patterns of Channel Reposting . . . 13
6.2 Virality . . . 13
6.3 Valency and Music . . . 15

7 Conclusions **16**

1 Introduction

Music is a ubiquitous force in our modern lives. It grounds us in the chaotic social landscape that we exist in. It is a fast-path communication mechanism [13]. Emotions and intentions that would require a long form writing to express can be packed into a few seconds of music. It is no wonder that music has risen to preeminence as a central

form of media, making up the media in online social media. Music can be combined with other forms of media to make something that would be short and simple into something highly evocative. Video, in particular, is often made more potent in combination with music. Different types of music can completely change the way a video is interpreted. Music is a powerful tool for film makers, to illicit different emotions from their intended audience; which music is used depends on what emotions the maker is attempting to convey.

Humans use a variety of shortcuts to more effectively present our viewpoints. Facts about reality are simplified and a shown only as what is important to get the point across. What particular parts a communicator chooses to emphasize or suppress is called framing. Frame analysis has been used in research as a way to analyze what biases exist in the way that information is presented [14]. Framing can be examined at a large scale, to give an interpretation of macro-level patterns of social expression [3]. Most of the research into framing has been using the modality of written text, however, frame analysis can also be used to interpret the content of multi-modal media.

Videos exist on the web at a massive scale as posts to social media. The way that these videos are shared and the type of music they are edited with provides a rich signal of human behavior and emotion. Keeping the underlying video context the same, different music can be used to drastically change the emotions that are induced. Social media sites are used by people from diverse backgrounds. The way videos are edited evokes different responses from different groups. Machine learning models can predict the difference in reaction between people of different political leanings [1]. For research intending to understand the characteristics of different factions on internet communities, large scale video data presents and interesting target to be examined using tools from data analysis.

Concepts from frame analysis can be adapted to apply to video-audio media. In this case, videos are taken to be videos capturing live scenes, or incidents. The underlying live scene itself is the raw unfettered information. Keeping the incident constant, different equivalence frames can be formed. The footage may be edited post-hoc, with certain parts being presented or withheld. This constitutes the framing of the footage. Also the footage can be edited with additional audio, such as music, which further modifies the presentation of the underlying information. Even more framing can be used; if the video is being shared on social media the title and comments and reactions all apply different filters over the base information. These all contribute to the framing of the scene, but the effects of them as a whole are difficult to elucidate. In order to simplify the complications of understanding all the possible frames in social media videos, this work focuses only on the interaction between the incident being filmed and the music.

2 Background

There are many social media platforms hosting videos edited to various music. Videos posted in the context of a wide range of events could be looked at from the perspective of frame analysis. I instead reduce the scope to media related to the war in Ukraine, and only examine videos from the Telegram messaging platform.

2.1 The War in Ukraine

In 2022 geopolitical tensions reached a critical juncture in Ukraine. In February the large scale invasion started. In Winter 2022 Russian forces were seemingly poised to capture the capital city Kyiv. The Summer and Autumn saw the fighting lines return close to where they were before the invasion, where they have remained since, as of the start of 2024. The invasion was accompanied by massive migration and refugee efforts. The events of the war are well documented. Owing to the prevalence of mobile cellphones and social media, many images and videos were shared. Many scenes of chaotic destruction of civilian infrastructure went viral. Videos of hitherto unheard of devastation get captured by civilians: Helicopters flying over the black sea [1], nighttime missile attacks, low flyovers of Russian jets. Footage released by military organizations also depicted the intensity of the war, and the new technologies being used. Many videos are captured by cheap consumer drones used for artillery spotting and from consumer drones jury-rigged to drop grenades. If the underlying footage is any indication of the intensity of resolve of the Ukrainian and Russian sides, the music added post-hoc serves to multiply the effect that the video posters intend to convey.

2.2 Telegram

Telegram is a social media platform founded in 2013 by Russian brothers Nikolai and Pavel Durov. The platform quickly rose to become one of the most popular social messaging apps in the world.

From a surface level Telegram appears like many other social messaging platforms. Users can register for the app using their phone number and send messages to contacts on their phone that have created an account with the

[1] `https://www.youtube.com/watch?v=zJHGNWdwIbw`

Channel Name	URL	Subscribers	Description
СМИ Россия - Украина	novosti_voinaa	3.4M	Russian news channel focused on news outside of the Moscow sphere, including war reporting.
Леонардо Дайвинчик	leoday	3.04M	Russian entertainment and funny memes
РИА Новости	rian_ru	3.0M	Russian news channel
Мир сегодня с ”Юрий Подоляка”	yurasumy	2.7M	Russian news channel focused on military and economic reporting
Труха Украина	truexanewsua	2.6M	Ukrainian news channel
Kadyrov_95	rkadyrov_95	2.1M	Channel of Ramzan Kadyrov, current head of the Chechen Rebublic in Russia
Mash	breakingmash	2.1M	Russian news outlet, mostly focused on domestic reporting
Рыбарь	rybar	1.2M	Russian military reporting, focusing on realtime war mapping and reporting of front line actions
Zelenskiy / Official	V_Zelenskiy_official	0.9M	Channel of Volodymyr Zelenskyy, President of Ukraine
Оркестр Вагнера | Wagner	orchestra_w	0.8M	Official channel of the Wagner mercenary group
GREY ZONE	grey_zone	0.6M	Russian telegram channel focused on Wagner

Table 1: Select popular Russian and Ukrainian language Telegram channels, subscriber counts as of Thursday November 2nd, 2023

messaging platform. Photos and videos and be sent at a higher quality than what is usually available over cellular messaging. Similarly, group chats are more convenient on Telegram than over default cellular messaging.

But personal chat features are not what sets Telegram apart from other messaging platforms. Telegram Channels are a unique feature of Telegram. They offer many-to-one communication; unlike a group chat, only one administrator is allowed to create posts. Channels can be marked as private or public, and administrators can specify whether or not they wish to allow other users the ability to comment. Channel owners can post videos and photographs as well as repost chats from other public channels. Telegram also supports a well documented and robust public API[2]. With the API, administrators of large channels are able to control and automate certain actions in the channel: posting, comment deletion, banning, and so on.

The combination of the large user base and the accessible developer API tools has allowed channels to rise to prominence as the premier outwards facing feature of the app. Especially in the Russian and Ukrainian sphere, most of the important political and media and news figures maintain active and popular Telegram channels. See Table 1 for a few select examples.

Telegram has also attracted fringe military and political groups to use channels as a base for distributing communications and propaganda. Perhaps as a result of the personal libertarian beliefs of Telegram's creator Pavel Durov, the platform does not conduct aggressive banning of many groups that would otherwise not be permitted on other social media platforms. In a short text post in Figure 1, Durov explains that while Telegram will remove content that is 'obviously harmful', he's hesitant to ban channels used by groups such as Hamas.

What information does Telegram make available to researchers? Telegram users using a free API key can collect certain data from public Telegram channels. Uploaded files and media, text posts, likes/reactions to posts, as well as public comments, view counts, and subscriber counts are all available for collection. Personal user-level information can not be easily aggregated and retrieved. As of 2023, Telegram is one of the most accessible mainstream social media platforms for researchers.

[2] `https://core.telegram.org/bots/api`

3 Related Work

3.1 Music Emotion Recognition

Music is a commonly used tool in film production to intensify the emotional experience of the film's video content [10]. The valence of a piece of media is defined as a measure of the emotional positivity. High valency corresponds to joyfulness, low valency corresponds to sorrowfulness.

Music Emotion Recognition (MER) is a field of study devoted to understanding capability of machine learning models to understand the relationships between music and emotion [4]. Core to MER is the obtainment of descriptive features from music such as valency, key, tempo and rythm.

Gómez-Cañón et al. [5] show that MER can be used to predict the induced emotions of people from different political leanings. They raise the idea of the potential capability, and danger, of using automated MER tools to influence users on social networks.

3.2 Research in the Telegram information sphere

Telegram has been used in previous social media research. The platform has liberal usage policies concerning API usage, allowing researchers to collect data on a massive scale. Morgia, Mei, and Mongardini [11] collected over 400 million messages in a massive dataset. They analyse meta patterns on the platform, looking at graph connectivity formed by channel reposting behaviors, and language topic modelling.

Hanley and Durumeric [6] collected 2.5 million telegram messages, mostly on Russia and Ukraine, then performing text topic analysis. They present their findings on patterns of information dissemination that exist on Telegram. They find that after the beginning of the February 2022 invasion the amount of messages spike considerably. Post-invasion, they find that the number of hyperlinks linking to western media sites declines. Their research indicates that Telegram is a major platform in the discussion of the invasion of Ukraine for Russian audiences.

3.3 Video Similarity

Different definitions of video similarity have been used by researchers. Revaud et al. [12] defined event retrieval, where two videos are defined as being similar if they are recorded at the same event in the same temporal time span. Kordopatis-Zilos et al. [8] define the concept of incident similarity. Incident retrieval has relaxed requirements for two videos to be considered similar. Under their definition, two videos are defined as similar if they both contain partial segments that depict footage occuring sometime in the same physical vicinity, in an overlapping time span. In this work I use incident retrieval as the definition of video similarity. This is for two reasons. Firstly, methods that perform well on incident retrieval benchmarks perform excellently on other video retrieval benchmarks. This holds true for benchmarks that have less strict definitions of video similarity than incident retrieval, and can been seen in the the work of Kordopatis-Zilos et al. [7]. Secondly, the relaxed definition of video similarity used by incident retrieval is useful for retrieving videos with many augmentations such as watermarking.

4 Ethical Data Collection: Privacy and Safety

Studying patterns of interaction on social media platforms means performing large scale data collection and analysis. Researchers need to be careful to avoid ethical violations that could harm privacy of platform users.

The data collected for this research was limited to the public Telegram channels involved in the analysis. These Telegram channels all are configured as public channels. Their posted messages can be viewed by any user on the internet, regardless of whether they are logged into Telegram. The content posted by these channels is by any expectation intended to be visible to a large audience. Redistribution of these posts is anticipated by the channels; many videos and images contain watermarks that reference that channel.

The fields that were collected, as well as the Channels that were included in the collection are described in Sections 5.1 and 5.2. None of the collected data contains information that could identify personal Telegram accounts that viewed or interacted with posts.

Notwithstanding the publicity of these channels, researchers still have the responsibility to be careful with handling channel data. Public channels are able to post footage of people that would not consent to it being shared if given the opportunity. This is important because of the high stakes in the Telegram political sphere. One administrator of the channel MoscowCalling was arrested by Russian police in August 2023 [3], on charges of

[3] `https://www.bbc.com/russian/articles/c72e4d3edkeo`

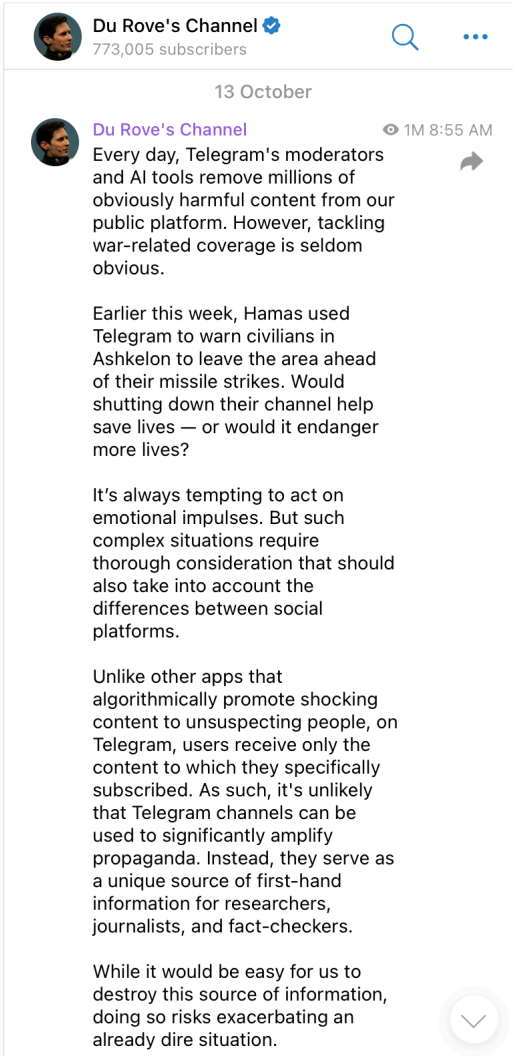

Figure 1: `https://t.me/durov/230` October 13th, 2023. Durov's discussion of Telegram content moderation with regards to military groups.

Channel Name	Predominant Language
colonel_cassad	Russian
combat_ftg	English
EurointegrationComUA	Ukrainian
hromadske_ua	Ukrainian
korrespondentnet	Ukrainian
liganet	Ukrainian
lost_generation_21	Russian
m0sc0wcalling	Russian
milchronicles	Russian
mozhemobyasnit	Russian
msgazdiev	Russian
opersvodki	Russian
OSINT_Ukraine_Aggregation	English
Pravda_Gerashchenko	Russian
rian_ru	Russian
RKadyrov_95	Russian
robert_magyar	Ukrainian
rybar	English
rysnya200	Ukrainian
ssigny	Russian
supernova_plus	Russian
svoboda_radio	Ukrainian
truexanewsua	Russian
Tsaplienko	Ukrainian
Ugolok_Sitha	Russian
ukrpravda_news	Ukrainian
vorposte	Russian
znua_live	Ukrainian

Table 2: Channels that were included in the analysis dataset

dissemination of fake information about the Russian army. There is a vast amount of information in the space of public Telegram channels, and it is the responsibility of researchers to mitigate the harmful risk of it's usage.

5 Methods

To establish equivalence frames means first obtaining a large number of videos sampled from real world social media. Clustering these videos requires an efficient video search pipeline. Then, analyzing differences in musical framing requires a system for identifying music tracks and their valency.

5.1 Channel Selection

Channels were qualified by several subjective measures. The channel should be actively posting content related to the war in Ukraine. It also should have at the very least tens of thousands of followers. Additionally, since this research is focused on video-music analysis, it should frequently post or repost videos.

Because of the large storage costs and computational costs of video retrieval, only a small number of channels were included. See Table 2 for a list of Telegram channels that were included in the analysis. Ten of these are official Telegram channels of Ukrainian news organizations. Two are of offical Russian news organizations, (rian_ru, mozhemobyasnit). Most of the channels advertised a affiliation with Russia or Ukraine in their channel description or channel name. Several of the channels are titled with disparaging names directed at the opposite side.

5.2 Telegram Data Collection

I used the Telethon [4] API to download post metadata and media from the set of selected channels. The final analysis totalled 188,217 video files, 346,175 images from 727,562 total scraped messages. Message metadata was stored in a database. Media files were stored on disk, referenced by file pointers in the database. All collection from Telegram was done using the free official API, using the Telethon python library.

5.3 Music Identification Pipeline

The music identification pipeline operates in two stages. In the first stage, for each video in the dataset, audio data is extracted and sent to the Shazam API for identification. Shazam returns a list of recognized song matches and their timestamps in the video. The timestamps and secondary matches get discarded. Only the metadata from the first returned match is saved in the database.

In a second stage, Shazam query results are converted to Spotify songs. Each recognized shazam song is queried using the Spotify API, searching for similar tracks based on the title and artist name. Results are filtered based on text similarity heuristics.

Because the music identification pipeline used two lossy stages, errors occurred frequently as discussed in Section 5.9.

5.4 Video Similarity Scoring

There were 188,217 videos downloaded from Telegram, but the number of unique incidents portrayed in these videos is much lower. In order reduce the raw video footage into some set of unique incidents, I obtained incident similarity scores between all pairs of videos. The number of pairs of videos scales quadratically with the number of videos. Since the number of videos is large, the speed of the search pipeline is very important. Distill and Select (DnS) [7] is a machine learning based method for retrieving videos based on incident similarity. It is a middle ground between performance and accuracy, and is also open-source. The authors released a model that was trained and evaluated on a diverse set of YouTube video footage from catastrophes. DnS was reported to be able to perform a single query over a dataset of two hundred thousand video data set in about five seconds. Considering those factors, DnS struck a good balance between accuracy and scalability for this project.

DnS uses an indexing pipeline that uses several models in order to compress the raw video bits into a compact and queryable format. For each video in the dataset, frames are extracted at one frame per two seconds. In practice I limit the total number of collected frames to 60, meaning only the first two minutes of each video are considered. The distill and select paper reported results on frames collected once per second, however I reduced this rate in order to shrink the compute and storage requirements. A ResNet50 convolutional neural network converts these frames into a sequence of 3D features. The pretrained `ResNet50_Weights.IMAGENET1K_V1.transforms` from torchvision was used. This model was to predict ImageNet classes. ResNet can extract features with some invariance to input resolution, thus different videos are embedded with different spacial resolutions. and stored with different spacial dimensions in an HDF5 dataset. The ResNet features are then passed through the DnS indexing networks, to produce course grained, fine grained, and selection network features. The indexed course grained and selection network features for the entire dataset total less than 500MB are are global features, meaning they are the same size for each video and can be represented in contiguous arrays. The indexed fine grain features, however, have different shapes depending on the spacial and temporal dimensions of the input videos. The fine grained features total roughly 80GB and are kept in a HDF5 dataset.

The querying process of DnS operates in a two stages: the fast but inaccurate course-grain (CG) stage, and the slow but accurate fine-grain (FG) stage. In the first stage the CG model calculates the similarity between the query video and dataset videos. Next, the selection network chooses which dataset videos should be re-evaluated by the fine-grained model in the second stage. The larger fine-grained model then calculates the similarity between the query video and some top proportion of the dataset videos as determined by the selection network. The second stage constitutes the bulk of the processing time, but is necessary because it provides a large boost in accuracy. The top kth percentile of the selection networks selected videos is a hyper parameter that can be adjusted; lowering k allows faster querying, at the cost of accuracy. The DnS paper recommends to setting k to the fifth percentile. In this work I use $k = 2.5$.

5.5 All-Pairs Video Incident Similarity Computed over a Web-Scale Dataset

The authors of DnS did not provide production code for using their model to compute all pairs similarity. But there are several tricks that can be used to ameliorate the expensive runtime costs of the retrieval pipeline. Queries

[4] `https://docs.telethon.dev/en/stable/`

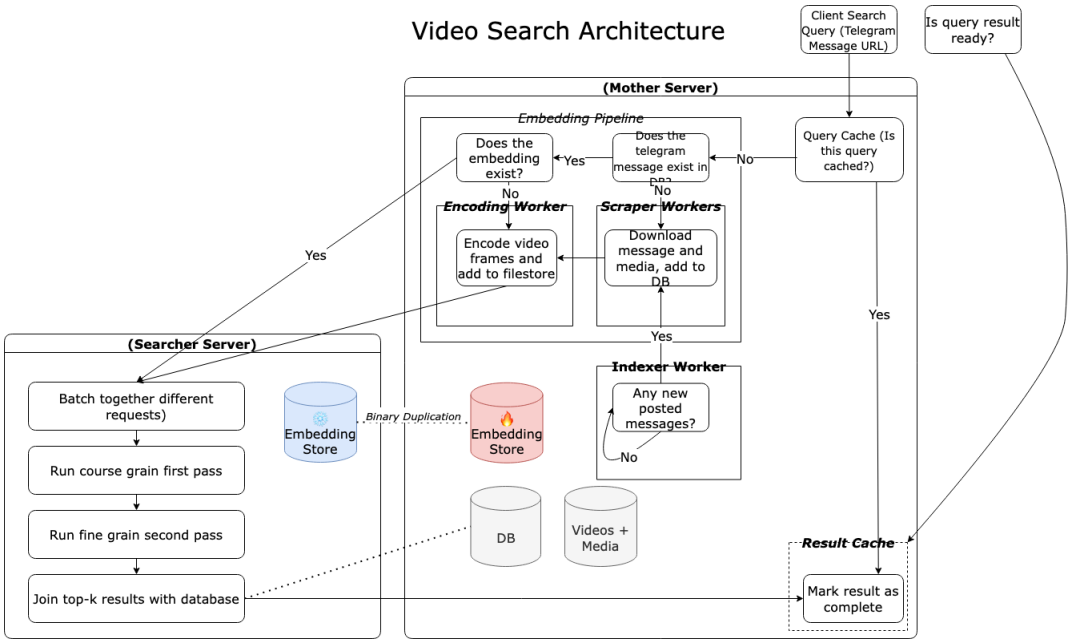

Figure 2: System diagram of the similarity pipeline.

can processed in batches, I use a batch size of 16. The CG and SN networks can be run in a single step, but the FG similarity computation should be split into minibatches to avoid excessive memory usage from batched matrix multiplication. Batching queries together comes with the additional benefit that some of the FG features selected by the SN can be de-duplicated. In practice I split the different steps in the querying and indexing pipeline into different tasks to be used with Celery and RabbitMQ. This allows the expensive FG similarity computation to be run in parallel on separate computers, keeping the core indexing and database management on one main server. The searches are distributed between workers by a load distributor. See Figure 2 for a diagram of the system design.

Before clustering videos, I obtained a pairwise similarity scores between all of the pairs of dataset videos. The desired result is a $\binom{n}{2}$ matrix video similarity scores. I simply query for each dataset video by inputing it into the video similarity pipeline, entering it's result into the corresponding row in the similarity matrix. This process is very time inefficient and expensive. Even using two gpu-equipped worker machines to support the querying task, the search process took 8 days.

5.6 Future Performance Enhancements

There are several improvements which could be made to reduce the processing time. One method which could drastically speed up similarity comparison would be a preprocessing step to cluster videos based on their selection network outputs. The selection network takes as input a pair of dataset video features, and ouputs a score determining the importance of sending this pair of videos to the expensive fine-grained model. A major bottleneck in querying performance is fetching fine-grain features from the disk. This bottleck can be eased by batching videos to be queried in a way that decreases the diversity of the fine-grained indices being selected. In other words, increase the intersection in the fine-grained features selected in the reranking process.

The FG model uses the Chamfer Distance to construct the similarity of all pairs of frame-level feature vectors. Allowing the Chamfer Distance metric to be computed using quantized feature vectors would significantly reduce the amount of data needed to be fetched and stored for the FG model.

A more powerful frame-level feature extractor could potentially allow for better retrieval accuracy. CLIP models have been trained that are more powerful than the ResNet model used in the DnS paper.

Kordopatis-Zilos et al. [9] proposes a refinement to DnS that allows for training and finetuning the retrieval model in an unsupervised manner. This shift to self-supervision presents a powerful paradigm for researchers. Using this paradigm, researchers can fine-tune the retrieval model on their own data without needing to create labels.

5.7 Clusters of Similar Videos

The result from the all pairs video search is an N by N matrix of similarity scores, where N is the number of videos in the dataset. These scores are asymmetric because of the small effect of batching on the search results.

I used the `sklearn` implementation of DBSCAN to produce clusters from these similarity scores. DBSCAN has a hyperparameter epsilon, which acts as a threshold to the distance between two data points. Using the elbow method, I choose an epsilon value of 0.20. Figure 4 shows the curve of number of clusters VS epsilon. I found that this parameter produced high quality clusters with minimal false positives, whilst still maintaining robustness to video editing.

5.8 Video Clustering Failure Points

Several of the largest clusters consisted of various cell phone shots of nighttime missile attacks. Nighttime scenes were difficult for the DnS pipeline to obtain similarity scores. The large clusters consisting of these videos were dropped by hand from the analysis. Nighttime scenes are a problem because the image embeddings fed to the DnS model are obtained with a ResNet model trained on ImageNet [2]. ImageNet collected images with the goal of increasing image diversity. The authors of ImageNet measured the diversity of a set of images by taking the averaged pixels values of a set and then measuring it's JPG filesize. A set of images all containing very dark pixels will have a low diversity score, according to the ImageNet metric. Very dark compressed images aren't common in the ImageNet dataset, which explains why this is a weak area of DnS. An example of this large cluster can be seen in Figure 6, where around 1,000 videos containing dark frames were grouped into a single cluster. I also observed this occurring for small clusters of videos, this can be seen in Figure 8. The large clusters of these dark videos are easily removed by hand. The smaller clusters are more difficult to remove, but as suggested by Section 5.10, they are probably unlikely to occur frequently.

Other issues could be interpreted to be associated with the vague definition of incident similarity. In the clustering step, each cluster is assumed to be associated with one and only one incident. However, videos may portray two or more incidents, simultaneously or consecutively. For example, in Figure 7 there are, arguably, two incidents being shown in one video. Because of the design of the clustering, videos can only be matched with strictly one incident.

5.9 Song Matching Failure Points

The song matching pipeline had frequent errors, which manifested as false positives, false negatives, and incorrect song assignments. Because of the opacity of both the Shazam and the Spotify services that were used to identify music, there is not an easy way of elucidating the exact reasons why these errors occurred.

In the previous experiment where I examined the false positive video clusters, I also looked at the assigned Spotify songs. False negatives are easy to identify, this occurs when the video has music and no Spotify song was assigned. In other cases Shazam would assign different song metadata to videos with identical file hashes. This suggests that the quality of the Shazam or Spotify recognitions was possibly impacted by rate limiting.

False positives appeared in more surprising ways. Sometimes a soundbyte was used independently in both a song, and a video. Even if a video doesn't contain any audio from the song, they both contain the same soundbyte and are matched by the Shazam algorithm. The voice of a woman speaking in sterile Ukrainian *"Увага, повітряна тривога"* ("Attention, air alert") - this soundbyte is an air alert warning all too familiar to residents of Kyiv. The audio was used by the Russian band Алоэ Вера (Aloe Vera) in their song *Новости* (*News*) [5], (see Figure 10). This song was falsely matched with a cluster of humorous meme videos containing the same sound byte: A video of a mother cat carrying her kitten by the scruff, retreating beneath a sofa.

The difficulties in the music recognition pipeline stress the need for robust valency scoring of audio. As mentioned in 5.3, some progress has been made in MER. Still, no open-source systems currently exist that could be used for large scale valency scoring over highly noisy audio files.

5.10 Manual Labelling

Because of the failure of the song matching pipeline, I needed to manually fix a portion of the data. I selected 339 clusters at random. I went through each cluster, playing videos simultaneously in a grid using the gridplayer app. I marked videos as false positive if they are obviously not captured from the same incident as the other videos in the cluster. False negatives are not easily discovered, and were not recorded. I also attempted to repair some of the song labels; in the case that one or more video in the cluster had the correct music, and other videos did not,

[5] *Новости*, Алоэ Вера `https://youtu.be/RxgKp13vp8Y?t=117`

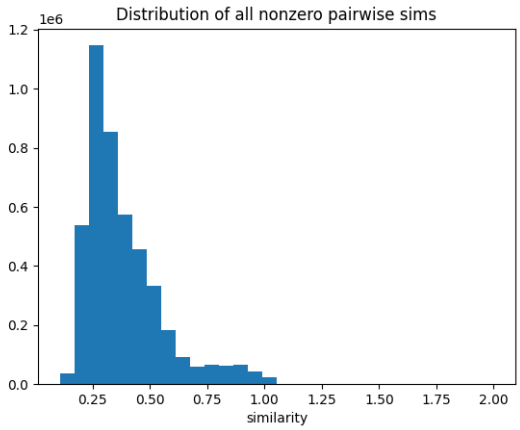

Figure 3: Similarity scores between all pairs of data set videos. Most scores are between 0.2 and 0.6.

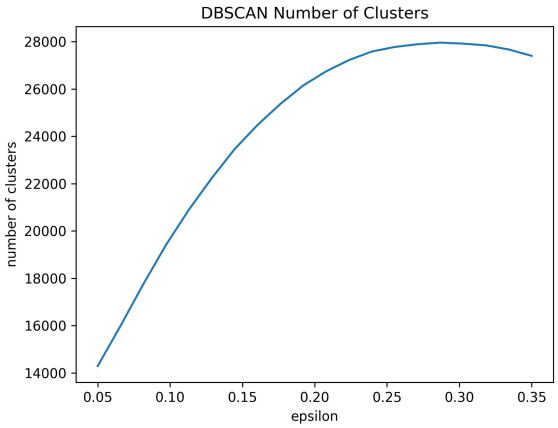

Figure 4: Increasing values of epislon cause larger numbers of total clusters, up to about 0.27, where the number of clusters starts to decrease.

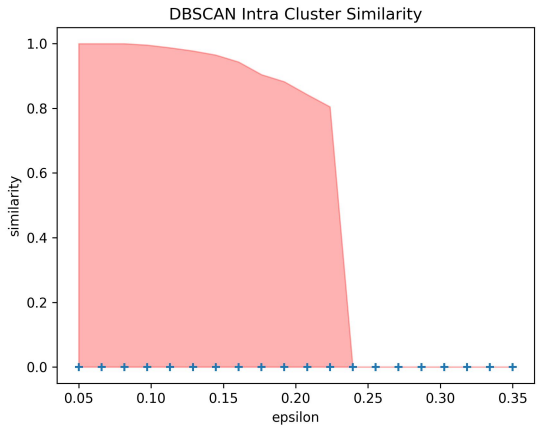

Figure 5: 95th percentile intra-cluster similarity scores. This highlights the asymmetry of the similarity scores, many clustered videos have zero similarity in one direction, but large similarity in the inverse direction. The blue marks indicate the median similarity, which surprisingly remains close to 0.0 for all epsilon values.

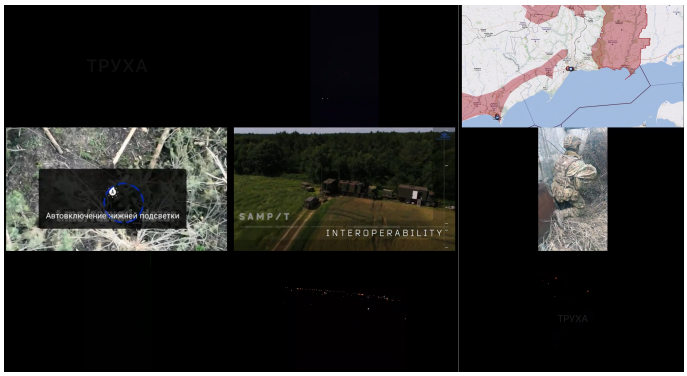

Figure 6: 1057 videos were assigned to this cluster. All of the videos had brief or extended scenes with very dark pixels.

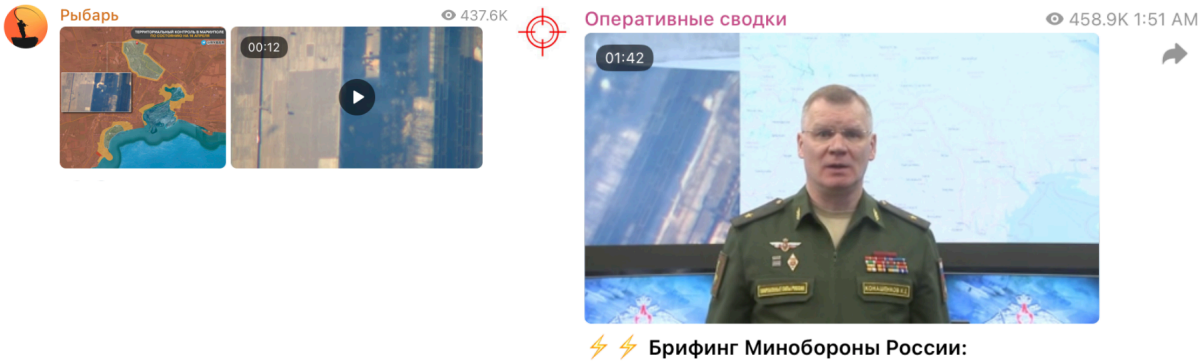

Figure 7: These two videos were recognized as similar. They both contain drone footage of a train yard. In the right video, the train yard footage can be seen playing on a screen in the background.

Figure 8: 8 videos were assigned to this cluster, all depicting night-time anti aircraft activity. These videos are not captured from the same incident and are false matches.

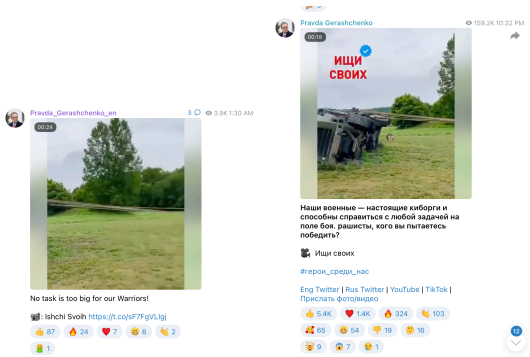

Figure 9: Two videos with identical file hashes, the first was correctly assigned the song Eye of the Tiger. In the second video the music wasn't recognized. `https://t.me/pravdaGerashchenko_en/18878`, `https://t.me/Pravda_Gerashchenko/42078`

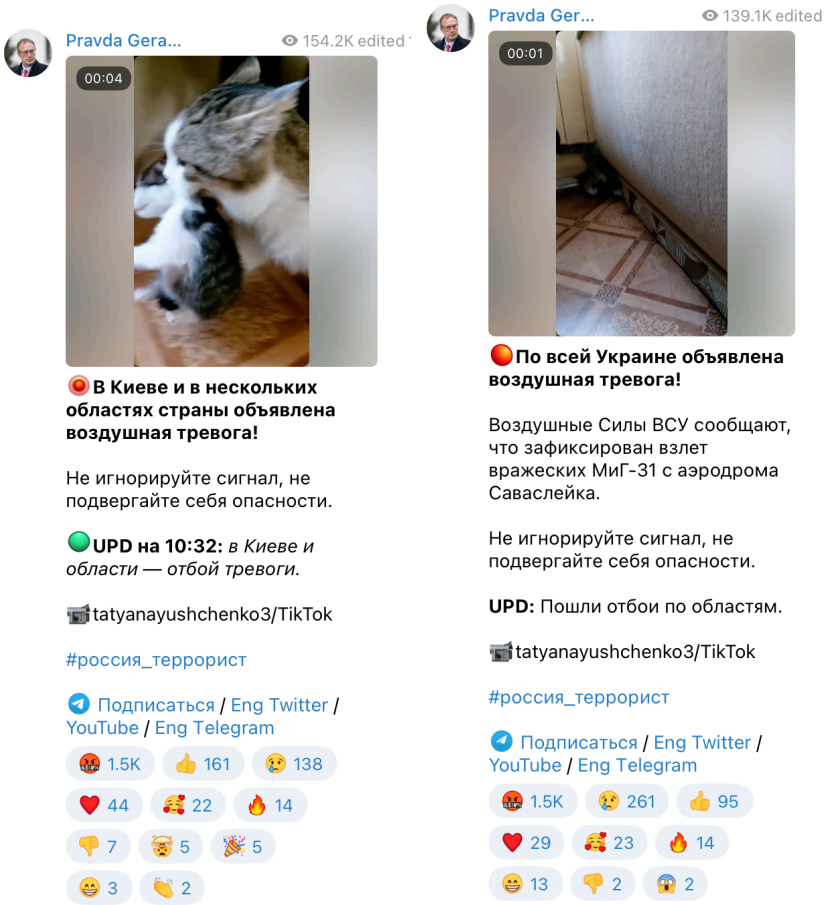

Figure 10: A cluster of videos that uses a soundbyte from an air raid warning. This cluster was matched with the song Новости by Алоэ Вера, which also uses the same soundbyte. Shown here: `https://t.me/pravda_Gerashchenko/77394` and `https://t.me/pravda_Gerashchenko/75564`

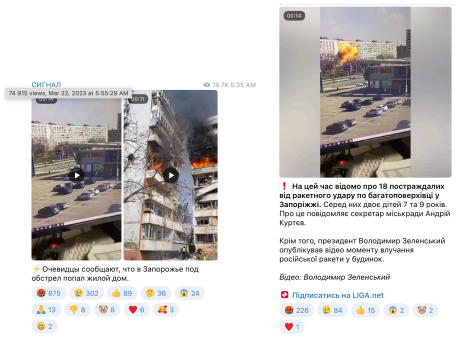

Figure 11: This cluster contains 10 videos, and shows a handheld phone recording of a desktop PC monitor playing back CCTV footage of a missile attack in Zaporizhzhia. In the background of the video there is a synth phone ringtone playing. Shazam assigned different music to all of the different reposts of this video, most of these being unrelated synth music. `https://t.me/ssigny/57276`, `https://t.me/liganet/25271`

I copied the song metadata from the correctly annotated video to the others. In the case that music metadata was falsely associated with a video that obviously had no music, I removed the music metadata. I did not fix songs that had failed to be recognized. Additionally, I identified 17 frequent classes of video content and assigned each cluster to a single class. The categories are shown in Table 3.

Out of the 339 clusters, only 4 clusters had at least one false positive video match. Out of the 4077 total videos, only 54 videos were false positives. This low rate of false positives suggests that most of the erroneous clusters are due to predictable issues with DnS such as the incapability to discriminate between videos with mostly dark pixels.

6 Results

6.1 Patterns of Channel Reposting

Different Telegram channels have different behaviors of reposting. Some channels are generally sources of information and post original videos. Others rampantly reupload content sourced from other channels. Between the Russian and Ukrainian channels, some channels stood out as reposting content from both sides. Others stood out as usually only reposting content captured from their side.

In particular, the Russian leaning channel MoscowCalling frequently occurs in clusters from footage captured by both sides. Sometimes MoscowCalling appears in clusters where Ukrainian channels are sharing videos with music intended to induce pro-Ukrainian sentiment. Curiously in these cases, the MoscowCalling videos will not have any accompanying music. One example of this is a video from March 18th, 2022 showing a fast paced montage of drone footage showing Ukrainian strikes against Russian positions. Ukrainian channels Pravda_Gerashchenko, supernova_plus, znua_live, truexanewsua, and unaffiliated vorposte show a version of the video that has the high energy techno Hip-Hop song Live Another Day by Kordhell. MoscowCalling on the other hand, posted a version of the video without any music. Another example of this can be seen in posts by different Ukrainian leaning channels, combat_ftg, rysnya200, and ukrpravda_news. This video depicts a montage of strikes against various Russian military assets, and is edited to VOID WHISPERS by ovg! & Watergun Collective. The Russian channel, MoscowCalling, posted this video twice, both times without any of the accompanying music that was featured in the Ukrainian channels, (1 and 2). Perhaps this behavior is explained by MoscowCalling avoiding propagating overtly pro Ukrainian news content.

Drone videos appear commonly when filtering clusters by number of different songs. This is due to drone videos not usually being released with accompanying audio recorded by the drone. The choice whether to use music and the type of music that is added depends strongly on the leaning of the Telegram channel. Oftentimes, channels sharing a drone video depicting a battlefield victory will choose to include music. Footage depicting battlefield losses, if shared at all, will usually not include music.

6.2 Virality

Not every video clip that is posted on Telegram is destined to become viral. There is a distribution over the quality of virality. Very rarely, a video will become viral and be shared widely not only across Telegram but also across various news platforms. Virality very closely depends on the incident that is portrayed in a video.

Name	Number of Clusters	Number of total videos	Description
at video	2	11	Videos captured of anti-tank missile control systems, or personal anti-tank weapons.
civ destruction	64	316	Personal recorded videos by civilians showing destroyed infrastructure or buildings.
civ misc	30	92	Personal recorded videos by civilians.
civ war	22	113	Personal recorded videos by civilians showing war combatants.
drone dropped grenade	10	36	Videos of explosive drones and drones with a detachable grenade payload.
drone landscape	13	58	Landscape footage captured by drones.
news misc	58	645	Videos that are captured first hand by news organizations.
news studio	6	500	Videos that are captured first hand by news organizations, inside a news studio.
other	9	41	Miscelaneous
podcast	4	367	Videos that accompany either an audio podcast, or formal video podcast.
political	39	1550	Conference footage or news footage of known political figures.
reunion	2	4	Video depicting soldiers returning home and greeting their families.
soldier personal video	51	199	Personal helmet camera or cell phone footage captured by a soldier.
soldier pro video	29	94	Studio quality edited footage of soldiers.

Table 3: Different content classes used for cluster labelling

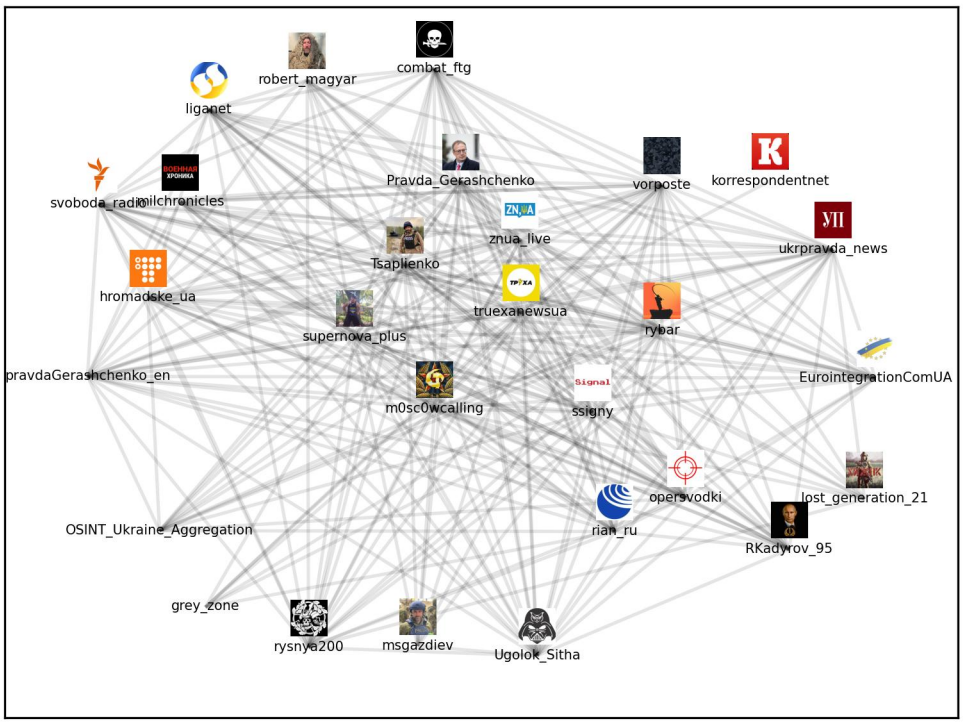

Figure 12: Cross-channel coocurrances visualized in a spring graph. Coocurances are defined as the number of pairwise occurrences of two channels in the same incident. This was calculated from the large scale unsampled data, before any manual labelling.

Channel moderators know which incidents have more impact and are likely to become viral. This is manifested in the choice of which incidents recieve additional post-hoc editing, and which don't. Channels budget their effort, editing videos that have a better ability to become viral. Figure 14 shows videos from the incident cluster of the CCTV footage of the 2022 missile attack on the Crimean bridge. Various Russian and Ukrainian channels reposted it, Prava Gerashchenko's channel warranted the video for more extensive editing, show in Figure 15.

The desire for virality can be seen in how aggressively channels market their own channel and it's logo using watermarking. A viral video can expand the popularity of a channel if it is posted with it's original watermark. Watermarks are extremely abundant. Very often an incident cluster will contain many separate watermarks, all vying for attention, flying across the screen and growing and shrinking.

6.3 Valency and Music

Spotify metadata comes with a predicted valence score. This score is a scalar value which represents how upbeat a song is. Taking the intra-cluster standard deviation provides a rough metric for how strongly the valence of music varies inside of a cluster, or in other words, how strong the misalignment of music valence is between different uploads of the same footage. Sorting the manually cleaned clusters by decreasing valence misalignment scores ranks the clusters by how misaligned the different songs are in the cluster.

The top few clusters, ranked in this way, have a pattern in how the channels share and edit the videos. One cluster shows a long distance drone shot of a Russian assault. The video is from Winter 2023, the men are standing in a system of trenches which are partially protected by a line of barren trees. Intermittent artillery fire is landing close to the trenches, throwing up large plumes of smoke and snow. The men leave their positions in retreat. Ukrainian leaning Supernova+ channel posted their version of the video with the song ВИТЯЗИ (Русские идут), "KNIGHTS (Russians are walking)", they also remove the ending logo screen of the 79th separate Airborne Assault Brigade of Ukraine, which is visible in the video from other channels. They instead add their own watermark into

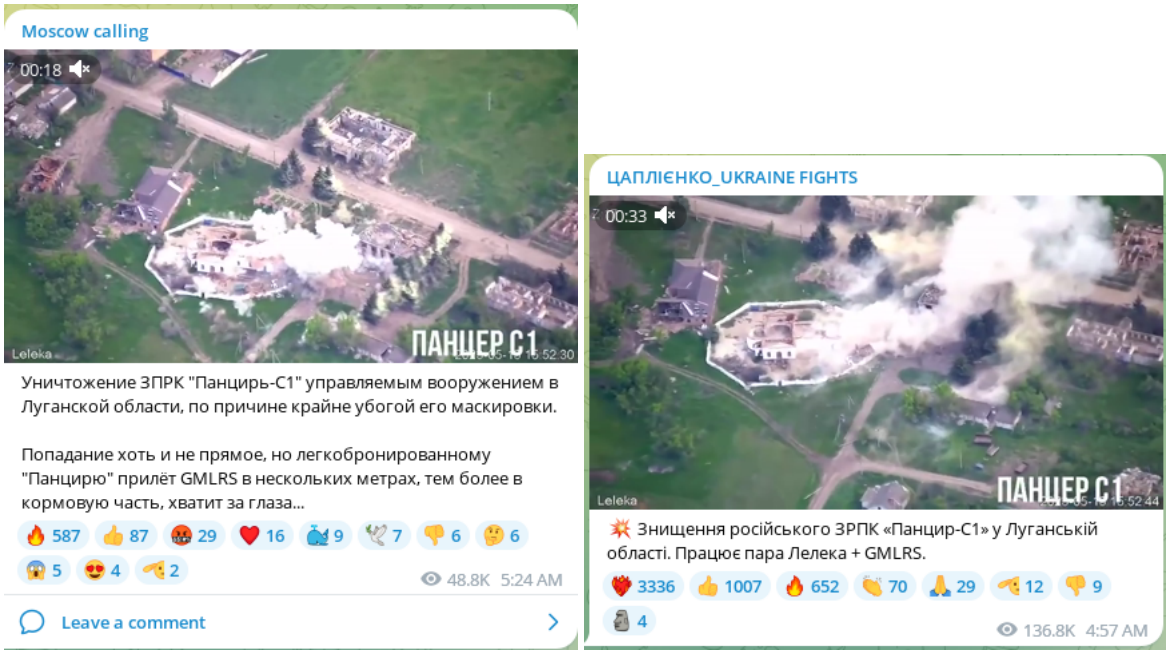

Figure 13: The left video, posted to the Russian channel m0sc0wcalling, had no music. The right video, posted to the Ukrainian channel, has the song $Feuerfrei$! by the German Industrial Metal band, Rammstein. Both videos show the destruction of a Russian anti-air vehicle. `https://t.me/m0sc0wcalling/24130`, `https://t.me/Tsaplienko/31798`

the video. Ukrainain leaning Pravda_Gerashchenko, znua_live, and rysnya200 post the same video, with a higher energy song, E.M.P. Eror (Smash Stereo Remix).

Perhaps if a better song recognition pipline was used, rare instances of factional valency misalignment could be found. Because videos and songs were reviewed manually, only the most common patterns made it into the analysis. No incident cluster was found to have vastly different music, edited to project polar opposite points. Instead, only a few clusters were found to contain different music at all, and when they did occur, most of the different tracks were in alignment. For example, Figure 17 shows two videos of Ukrainian rocket artillery firing in a muddy field. They don't have the same song, but both songs are high energy techo dance, and emphasize corresponding feelings. This pattern was characteristic of videos showing Ukrainian forces at work; different, but similar songs are edited in. Othertimes, more nuanced differences between music occurs in incident clusters. Figure 18 has two videos of a BTR combat vehicle, one post uses "Where are you from?", the other uses the X-Files theme song. These different edits aren't in opposition with each other, they both highlight the classification of the vehicle.

7 Conclusions

Looking at video sharing from the perspective of framing provides unique insights into the belief systems of factions. Maintaining equivalence frames, and analysing different musical framing methods, reveals the stereotypes and assumptions that are made in communicating inside of a faction. Videos portraying certain incidents were posted by both sides. The CCTV footage of the strike of the Crimean bridge is one such incident. This incident was interpreted in different ways by both factions which manifested in the editing of video. Some classes of incident seemed to not exhibit such extreme polarity between factions. Drone footage of battlefield scenes did have different music edited in between different sides, but there was not an large difference in the valency between both sides. The commonly observed pattern was the Ukrainian victories being edited and shared on Ukrainian channels with music in triumph of battlefield victories. These videos, if shared at all on Russian channels, usually have no music.

Video footage of incidents can be clustered by existing methods to form robust equivalence frames. Future research could benefit from larger datasets, and better music valency recognition techniques.

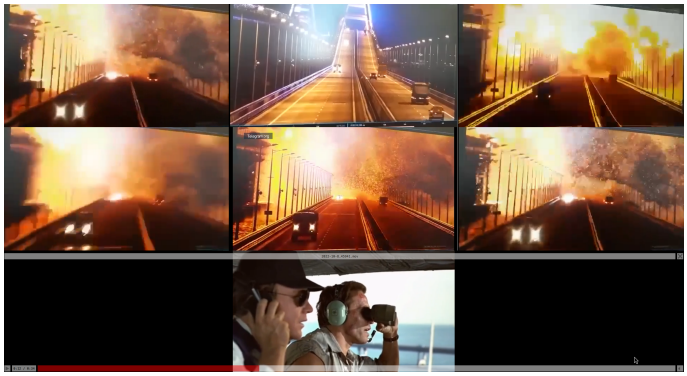

Figure 14: This video of the 2022 explosion of the Crimean Bridge was shared and edited with various different spacial cropping, and time-scale cropping.

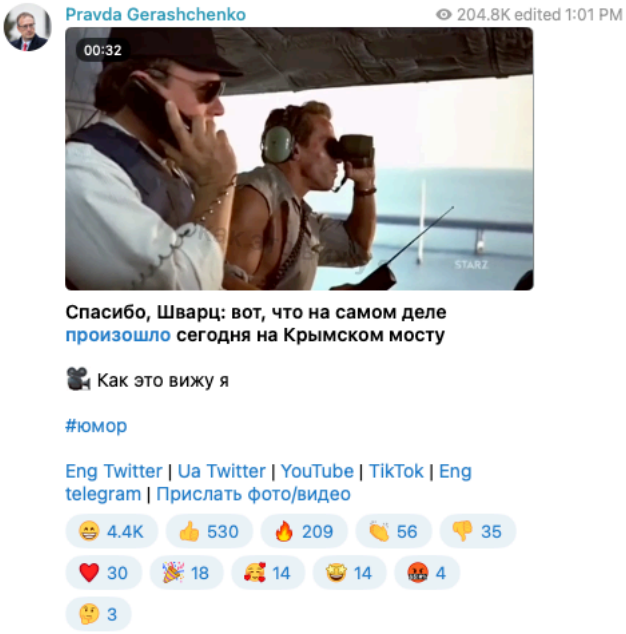

Figure 15: CCTV footage of the Crimean Bridge explosion, edited into a scene from True Lies where Arnold Schwarzenegger relays a command from a helicopter to destroy a bridge. `https://t.me/Pravda_Gerashchenko/45941`

References

[1] Phoebe Chua et al. *Predicting emotion from music videos: exploring the relative contribution of visual and auditory information to affective responses*. 2022. arXiv: `2202.10453 [cs.CV]`.

[2] Jia Deng et al. "ImageNet: A large-scale hierarchical image database." In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: `10.1109/CVPR.2009.5206848`.

[3] David Garcia et al. *Social media emotion macroscopes reflect emotional experiences in society at large*. 2021. arXiv: `2107.13236 [cs.SI]`.

[4] Juan Sebastián Gómez-Cañón et al. "Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications." In: *IEEE Signal Processing Magazine* 38.6 (2021), pp. 106–114. DOI: `10.1109/MSP.2021.3106232`.

[5] Juan Sebastián Gómez-Cañón et al. *Personalized musically induced emotions of not-so-popular Colombian music*. 2021. arXiv: `2112.04975 [cs.SD]`.

[6] Hans W. A. Hanley and Zakir Durumeric. *Partial Mobilization: Tracking Multilingual Information Flows Amongst Russian Media Outlets and Telegram*. 2023. arXiv: `2301.10856 [cs.CY]`.

[7] Giorgos Kordopatis-Zilos et al. “DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval.” In: *International Journal of Computer Vision* 130.10 (Aug. 2022), pp. 2385–2407. DOI: 10.1007/s11263-022-01651-3. URL: https://doi.org/10.1007%2Fs11263-022-01651-3.

[8] Giorgos Kordopatis-Zilos et al. “FIVR: Fine-grained Incident Video Retrieval.” In: *CoRR* abs/1809.04094 (2018). arXiv: 1809.04094. URL: http://arxiv.org/abs/1809.04094.

[9] Giorgos Kordopatis-Zilos et al. *Self-Supervised Video Similarity Learning*. 2023. arXiv: 2304.03378 [cs.CV].

[10] Chia-Wei Li, Tzu-Han Cheng, and Chen-Gia Tsai. “Music enhances activity in the hypothalamus, brainstem, and anterior cerebellum during script-driven imagery of affective scenes.” In: *Neuropsychologia* 133 (2019), p. 107073.

[11] Massimo La Morgia, Alessandro Mei, and Alberto Maria Mongardini. *TGDataset: a Collection of Over One Hundred Thousand Telegram Channels*. 2023. arXiv: 2303.05345 [cs.CY].

[12] Jérôme Revaud et al. “Event retrieval in large video collections with circulant temporal encoding.” In: *CVPR 2013 - International Conference on Computer Vision and Pattern Recognition*. Portland, United States: IEEE, June 2013, pp. 2459–2466. DOI: 10.1109/CVPR.2013.318. URL: https://inria.hal.science/hal-00801714.

[13] Jay Schulkin and Greta B. Raglan. “The evolution of music and human social capability.” In: *Frontiers in Neuroscience* 8 (2014). ISSN: 1662-453X. DOI: 10.3389/fnins.2014.00292. URL: https://www.frontiersin.org/articles/10.3389/fnins.2014.00292.

[14] *WebSci ’20: Proceedings of the 12th ACM Conference on Web Science*. Southampton, United Kingdom: Association for Computing Machinery, 2020. ISBN: 9781450379892.

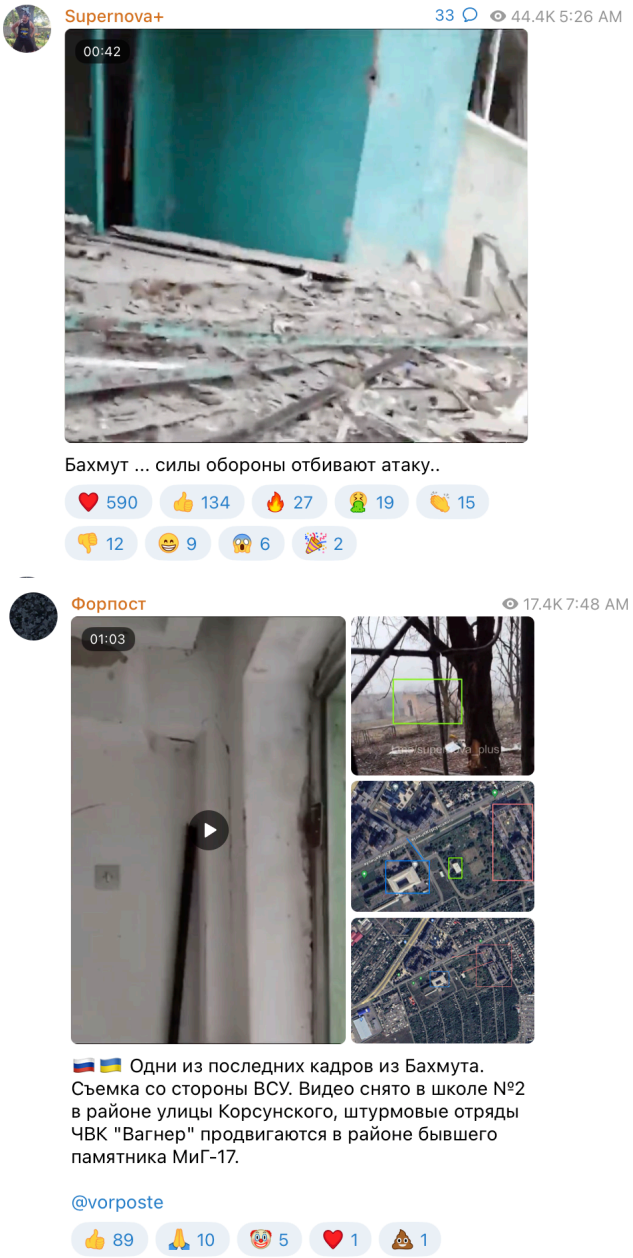

Figure 16: Helmet camera footage from a Ukrainian soldier in Bakhmut. The left video from the Ukrainian channel supernova_plus is edited to the rap song Dirty Money by Memphis Cult. The video from the Russian leaning channel Vorposte has the same supernova_plus watermark, but no audio. `https://t.me/supernova_plus/18525`, `https://t.me/vorposte/37424`

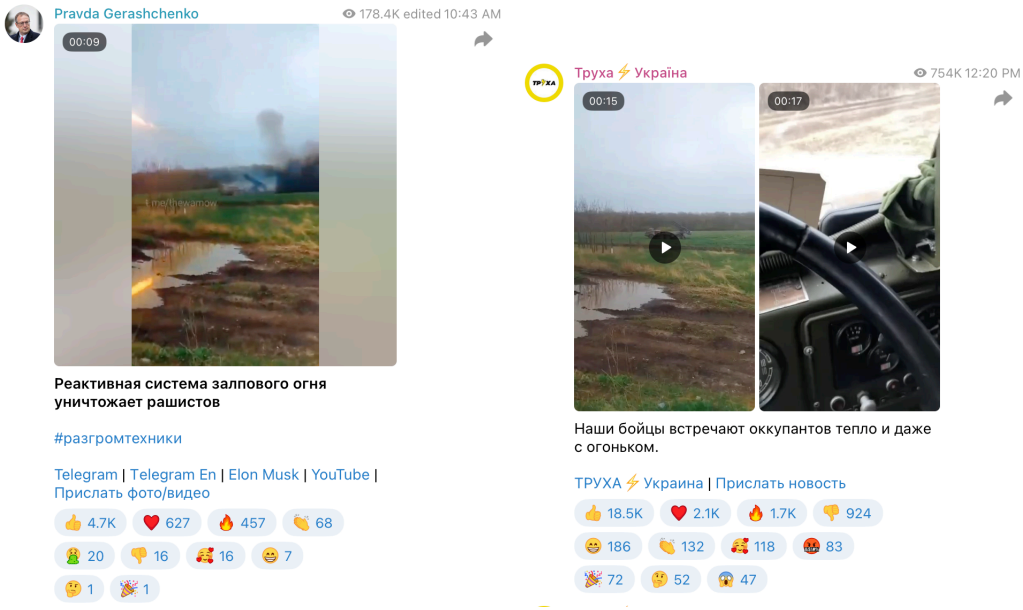

Figure 17: Same incident videos. The left video has the song Stay Mad by IV JAY, the left has the song Зомбі by TARAS KEEN. These songs have similar valence. `t.me/pravda_gerashchenko/15326`, and `t.me/truexanewsua/43992`

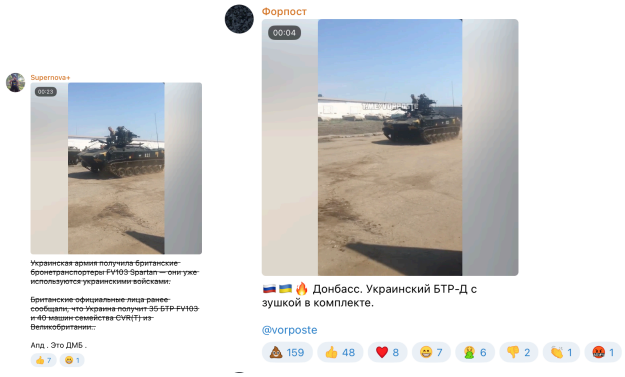

Figure 18: The left video uses the main theme song from the X-Files, the right video uses 'Where are you from' hard bass. `https://t.me/supernova_plus/5513` and `https://t.me/vorposte/20511`