Characterization of Social Video

Jeffrey R. Ostrowski and Nabil J. Sarhan

Wayne State Media Research Lab and Dept. of Elec. & Computer Engineering

Wayne State University

Detroit, MI 48202 USA

ABSTRACT

The popularity of social media has grown dramatically over the World Wide Web. In this paper, we analyze the video popularity distribution of well-known social video websites (YouTube, Google Video, and the AOL Truveo Video Search engine) and characterize their workload. We identify trends in the categories, lengths, and formats of those videos, as well as characterize the evolution of those videos over time. We further provide an extensive analysis and comparison of video content amongst the main regions of the world.

Keywords: Social media, video streaming, workload characterization

1. INTRODUCTION

Interest in social media has grown dramatically across the Internet and continues to evolve. Instant messaging, online gaming, photo sharing, video sharing, social networking, and other online social interaction have all been ways that technology has connected people and increased avenues of communication. Photo sharing websites such as Flickr have recently gained popularity since their launch in the last several years,[1] while massively multiplayer online role-playing games such as World of Warcraft have steadily increased revenues.[2]

People have embraced video sharing most recently to entertain, inform, and teach. Webcams, cell phones, cameras, and a multitude of portable devices have enabled people to capture information in the form of video and share it with the world instantaneously. This is further underscored by the U.S. release of the iPhone, for example, in June 2007 which included YouTube as a built-in, marketed feature[3] - not just a potential capability. Instead of watching a favorite TV channel for a limited set of weekly programming, users of sites such as YouTube and Google Video are able to view as well as post thousands of videos for public view at no charge. These websites, and others like Metacafe, Yahoo! Video, and MySpace, have grown drastically in popularity. YouTube, for example, is now the third most-viewed website in the world according to Alexa Web Ranking.[4]

This paper investigates online social video content with a global perspective. In particular, it characterizes videos from YouTube, Google Video, and the Truveo search engine in terms of popularity distribution in daily, weekly, monthly, and yearly contexts over a seven-month period, more than any other social video study of which we are aware. It also studies the correlation between video categories and popularity rankings and how the video access changes with the day of the week and the time of the day over an eight week period. In addition, this paper examines how the video access evolves over time. Moreover, it analyzes channel and video format encoding popularities. Furthermore, it studies the social aspects and how video category popularities vary depending on the users' world region. We approach these characterizations with the widest lens possible, not focusing only on a specific user group, category, or region, but rather including the entire compilation of users who participate in social video programming each day. The characterization of social media helps in understanding user behavior and the regional/social influence, as well as the design of delivery systems.

Very few papers have focused on characterizing social video. The study [5] focused on YouTube video clip popularity, payload, and data rates to determine that client-based caching can reduce network traffic and allow faster access to YouTube videos. It sought to optimize network traffic on a local network, which is different from our global approach, and did not analyze other social media sites. It also did not use popularity metrics directly from YouTube, which we do through the use of the available YouTube data services. The study [6] presented the popularity characteristics of YouTube and Daum,

For further author information - Email: {jeff,nabil}@wayne.edu, Telephone: 001-313-577-2680.

including the exhibition of power-law in the evolution of popularity. The data in that study were limited to "Entertainment" and "Science & Technology" videos over a period of several days. In contrast, we consider all categories and analyze YouTube popularity for a period of seven months. A different user-level approach to analysis of YouTube was taken in [7]. That study compared YouTube data with traditional website user sessions using duration, inter-transaction times, and content types, among other methods. User data were taken from users on a local area network as well, not characterizing all YouTube users as done in our study. In [8], a geographical approach to YouTube characterization was carried out in Latin America to identify the locality space for that region in comparison to the United States and the rest of the world. In our approach, we take a global view by presenting Google Video data from each world region. We further study the types of videos that are watched, not only in Latin America but all regions, to gain insight into categorical trends.

The main findings can be summarized as follows. (1) The popularity of social videos on YouTube and Truveo search engine data is characterized by a Zipf-like distribution, but with two different values of the skew parameter: $t = 0.30$ for daily popularity distributions and $t \approx 0.50$ for weekly, monthly, and yearly popularity distributions. This difference in t has not been previously been uncovered by any other study. The stretched exponential distribution suggested by a notable recent study[9] for characterizing the popularity of online videos does not accurately characterize online social video. (2) Music, comedic, and entertainment videos are the the most popular among all categories. (3) Most social videos published are less than five minutes in length, with the most popular videos averaging 3.5 to 4 minutes in length. (4) Video access frequency over time is characterized by an oscillating curve as popularity varies per day of the week. (5) Fridays and Saturdays from 9pm - 1pm UTC are the most popular times for users to access social videos. (6) YouTube is, by far, the most popular channel for accessing online social video, and the Flash video format is now available for over 80% of all videos published. (7) The category of music dominates social video popularity in Latin America and the Carribean, Northern America, and Western Europe. Geographical data we uncover is a starting point for understanding regional influence on social video selection.

The rest of the paper is organized as follows. In Section 2, we provide an overview of previous related work. Section 3 provides our evaluation methodology. In Section 4, we present and analyze the data in terms of popularity distribution, categories, days of the week and evolution, length, category comparisons, and social aspects. Section 5 concludes.

2. RELATIONSHIP TO PRIOR WORK

Many studies have characterized online and streaming video trends; however, few have focused on social video. We compare our work to both social video studies, as well as online video studies in general, in the following section.

The frequency of user access to videos is one of the main characteristics. In general, studies have either supported or opposed the notion that user access to online videos follows a Zipf-like distribution. In Zipf-like distribution, the probability of accessing video with rank v_r is

$$P(v_r) \propto 1/(v_r^{1-t}). \tag{1}$$

When analyzing social videos, study [8] presented data showing that the number of views when compared to rank did not exhibit Zipf's law for the data subset collected, in contrast with our findings. Further, the difference between that study and ours is that it took an aggregated snapshot of YouTube data over a period of eleven days. In our study, we collected data for the most popular videos each day for over seven months, allowing us to take a consistent view count average over a much longer period of time. Study [6] indicated adherence to the Zipf-like distribution; however, we believe its inclusion of only two categories ("Entertainment" and "Science & Technology") over several days is limited. In our work, we show how videos of each available category vary in popularity within a seven-month period.

Let us now compare our study with online video studies. Study [10] attempted to disprove that user accesses to online video followed the Zipf-like law. The environment for the study was a university network, with videos types characterized by both general entertainment videos (e.g., movies) and class lectures. The study distributed videos over a high-bandwidth network, removing network latency as a major factor for user behavior. Another study[11] characterized streaming media workload of streaming servers, including stream merging, server popularity, and session characteristics. That study analyzed client-based streaming media at the university and compared its characteristics with more traditional Web pages and HTTP protocols. It determined that user access frequency did follow a Zipf-like distribution, with $t = 0.47$, for media access from the university to the public Internet. Additional work[12] contributed to the debate when extensive client workloads were examined on two media servers at the University of Wisconsin-Madison and the University of Saskatchewan for classroom lecture videos and other course content. That study showed that user access frequencies on each server were

modeled by the concatenation of two Zipf-like distributions. Taking the analysis to the corporate workplace, paper [13] analyzed enterprise media server workloads for Hewlett-Packard Corporation in 2002 and correlated Zipf-like distribution with monthly video popularity metrics. More recently, the study [9] suggests that a stretched exponential distribution is best to characterize video popularity. In our study, we show, however, that the Zipf-like distribution characterizes the popularity of the top YouTube and Truveo Search videos.

One characteristic of social video that has not been widely studied is category. As we mentioned previously, study [6] limited its scope in comparison to a full categorical analysis of social video. Even though study [8] stated that it had collected video length data, it did not present that data for analysis. While its data were not specifically for social video, the study [10] analyzed distinctions between educational lectures and movies. It found that access to movies, for example, tends to be evenly distributed over time, while access to educational videos exhibits very high access rates over small periods. In our study, we take categorical analysis further by presenting results for all social video categories provided by YouTube, Truveo Search, and Google Video data.

Social video access evolution has recently been studied in [7] over a period of a day and in [6] over a period of a few days, which are much shorter than our eight-week analysis of access evolution. Comparing our study to previous online video analysis, the study [10] showed a cyclic pattern of access for videos which increases over time. Evolution analysis in [13] defined the life duration of a media file to be the time between the first and last accesses of the file in a given workload. That study further showed that for an enterprise media server, more than 50% of media file accesses occur in the first week of the file's existence. The paper [9] stated that 50% of requests were for objects older than 150 days. Our study focuses on social video data and takes a different approach, first identifying the most popular videos on social video websites and then tracking their progress over time by looking for cyclic patterns and life duration characteristics.

Length, another common video characteristic, is an important metric to understand when determining cache and bandwidth for serving social videos. The study [6] only mentioned two average lengths from its analysis: 30 seconds for the Daum Commercial Film category and 203 seconds for the Daum Music Video category. It did not correlate this data to any further analysis and did not show trends. Study [8] stated that 80% of the videos recorded in its analysis were less than five minutes in length, again showing no further trends as we do in our analysis. Looking toward a more in-depth online video length analysis, study [11] showed that the most common media streams were 3.5 to 4.5 minutes in length, leading to the conclusion that clients have a strong preference for viewing short multimedia streams. The study [13] further concluded that even though a proportional number of videos are accessed in an enterprise over the spectrum of video lengths, the duration for a majority of the accesses ($77\% - 79\%$) is less than 10 minutes in length. In our further study, we extend length distribution studies by presenting data for YouTube and videos indexed by Truveo Search, including video length distribution data by category.

We further contend that a geographical approach is necessary in modern social video characterization. In paper [8], a geographical approach to YouTube characterization was carried out in Latin America to identify the locality space for that region in comparison to the United States and the rest of the world. That study is an attempt to understand geographical factors in video services and how regional differences influences social video behavior. It is the first to recognize that social video trends may indeed be influenced by social aspects practiced within a given region. We expand this approach in our study, as we present social video data from each world region. We further study the types of videos that are watched in each of these regions to gain insight into categorical trends.

Building on the above work, we aim to characterize the video popularity distribution for social video websites, the types of videos which are accessed, how videos are accessed over a period of time, length metrics, and geographical influences. The summary of these aspects will help to provide insight into social video patterns.

3. EVALUATION METHODOLOGY

We collected data from two social video websites and one video search engine: YouTube, Google Video, and the AOL-owned Truveo Video Search.[14] We created scripts using the Ruby programming language, including scripts for the Google and YouTube developer interfaces.[15,16] Data were closely monitored and changes to the scripts were made to accommodate Google and YouTube website modifications over the collection period.

YouTube data was collected for a period of twenty-nine weeks. Daily, weekly, monthly, and yearly Top 100 video feeds from YouTube were successfully downloaded each day, including video ID, title, duration, view count, category, and URL. We then took a snapshot of the top 100 most viewed YouTube videos of all time and tracked their popularity evolution

over an eight-week period. We collected video IDs, titles, durations, view counts, category, and URL information for these videos as well twice a day at 9 PM Coordinated Universal Time (UTC) and 1 PM UTC. Once the data collection was complete, we calculated the daily view count for each video in our top 100 list, as well as the view count statistics for the periods of 1PM to 9PM UTC and 9PM to 1PM UTC each day.

Expanding our data collection, we then used data from the Truveo Video Search engine to understand the effect of a much larger set of online social videos. The top 1000 videos of all time were collected for a period of 29 weeks. Video popularity, category, length, and format metadata were extracted for these videos and normalized.

Lastly, we took a snapshot of the top one hundred most popular daily videos from Google Video. Not only did we collect the top 100 videos for the world ("All Countries & Regions") but also we captured the top 100 videos for each country and region for which Google Video provided a top one hundred list (Australia, Finland, Peru, etc.). After collecting the top one hundred videos and their associated rank in the top one hundred list for each country and region, we then proceeded to find the associated category of those videos. As Google Video incorporates YouTube, Google, and other online video websites into its lists, we had to come up with an intelligent strategy to get the category data. First, we parsed the top hundred list to determine if each video listed was a YouTube video. We did this by making a call directly to the URL of the Google Video ID, which soon informed us after some HTML parsing whether or not this was a YouTube video. We then were able to search and cross-reference the listed Google Video identified with the YouTube identifier that was parsed. From this YouTube identifier, we were able to appropriately search YouTube for the category information per our previous interface addition. Since the category information for Google videos was not readily posted on Google's website, we needed to find a way to get this information.

The Google Advanced Video Search engine allowed us to query videos for a particular category using the Google identifier. In order to determine whether a video was in one of the thirty-nine Google Video categories, however, we would have to search up to thirty-nine categories using the search engine until we found the category to which the video belonged. This could feasibly equate to more than 100,000 Google searches for all of the downloaded videos, over 175,000 searches for the worst case scenario. At the end of the automated exercise, we had only a handful of videos for which we needed to look up the category manually.

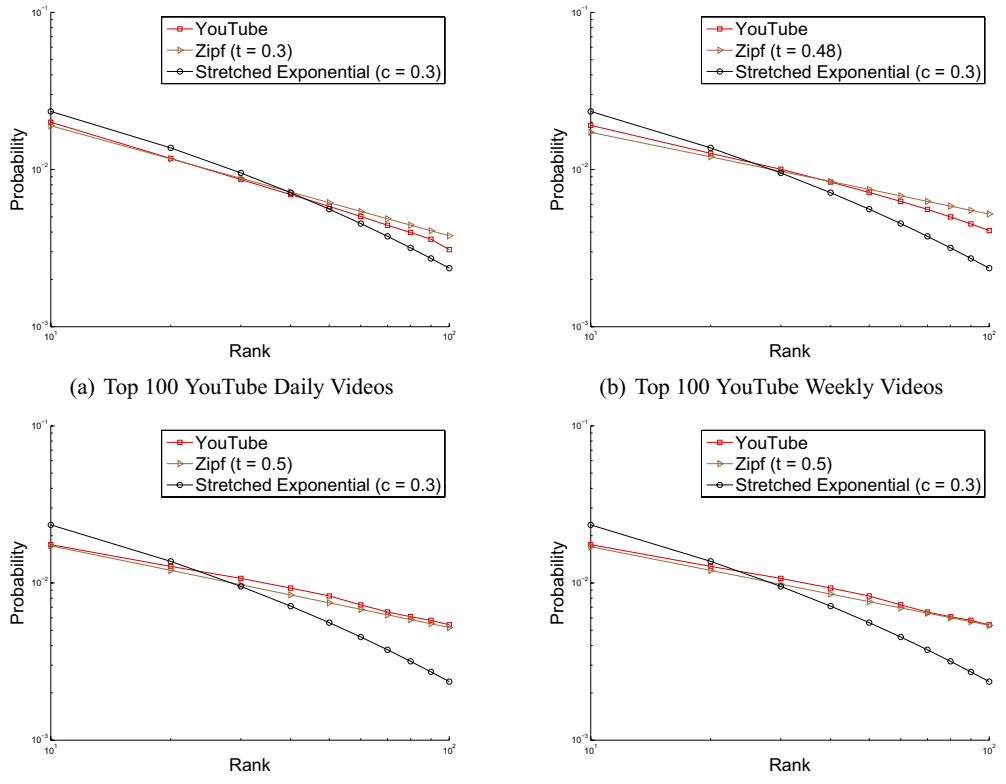
The retrieved video lists now needed to be related to one another based on country. In order to account for these geographic differences, we acquired the regional groupings of countries from the U.S. Census Bureau (Asia excluding Near East, Commonwealth of Independent States, Eastern Europe, Latin America and the Caribbean, Near East, Northern America, Oceania, Sub-Saharan Africa) and weighted the results we collected for each country based on population for the region from the U.S. Census Bureau International Data Base.[17] If Country X had double the population of Country Y, for example, that category information for a particular rank within the Country X data would then be weighted double in comparison to the category information for a particular rank for Country Y. This normalized the results between countries so that the population of the region was better represented. Our next issue, however, was to figure out how to ensure rank data was accurately represented when merged. In order to maintain the rank contribution, we also weighted each value further by the probability density function for each particular rank.

4. DATA PRESENTATION AND ANALYSIS

4.1 Popularity Distribution

In accordance with previous studies, we seek to characterize video popularity with the YouTube data that we have collected. Figure 1 shows the probability density function of video popularity in terms of access frequency (number of views) for the Top 100 YouTube Videos lists (daily, weekly, monthly, and yearly). As can be seen by the figures, the distribution for all popularity periods follow Zipf-like curves. Daily videos show very close similarity to the Zipf distribution with $t = 0.30$, while all other popularity periods (weekly, monthly, and yearly) show t near or equivalent to 0.50. This difference in t has not been previously been uncovered by any other study. In study [11], a week's worth of access frequency data showed that media requests over the public Internet produced Zipf-like results as well with $t = 0.47$, very close to our findings for weekly, monthly, and yearly distributions. Our findings show, however, that Zipf-like distributions for YouTube access frequency may be different based on the specified reference period. Daily access frequency for the top 100 videos must be characterized differently than weekly, monthly, and yearly videos. The aforementioned results demonstrate that daily accesses are highly skewed, with the top 20 videos accounting for over 40% of the total accesses. Monthly and yearly accesses are less skewed, with the top 20 videos accounting for 34% of the total accesses. Hence, we conclude that a

popular video in terms of daily access will have a higher probability of being accessed more frequently amongst its peers compared to a popular video in terms of monthly access (or beyond).



(a) Top 100 YouTube Daily Videos

(b) Top 100 YouTube Weekly Videos

(c) Top 100 YouTube Monthly Videos

(d) Top 100 YouTube Yearly Videos

Figure 1. Video Probability Distributions [Average over 29-Week Period]

One recent study [9] attempted to show that video access frequency is instead characterized by a stretched exponential distribution. That paper analyzed a host of different web and P2P data sources between 1998 and 2006 and demonstrated that sixteen analyzed workloads can be characterized by the equation

$$P(v_r) \propto e^{-t/{t_0}^c}, \tag{2}$$

where t_0 is the characteristic relaxation time (a constant), and c is a number between 0 and 1. As can be seen from the data plotted on a log scale in Figure 1, the recent YouTube data does not follow stretched exponential tendencies. The closest margin of error between the data and the stretched exponential form is exhibited in Figure 1(a), yet still a Zipf-like curve with $t = 0.30$ is a better fit for the data. We conclude that YouTube access frequency over daily, weekly, monthly, and yearly periods is more readily characterized by a Zipf-like curve and does not follow the stretched exponential distribution suggested in study [9]. Our further analysis reveals that smaller video populations (such as the top 50 videos) follow a similar trend and thus not shown to save space.

In order to further prove our findings with the YouTube data, we look toward a larger video population to understand if Zipf-like behavior is maintained. Figure 2 shows the popularity distribution of the Top 1000 Videos (All Time). The videos follow a Zipfian distribution with $t = 0.30$, different than the $t = 0.50$ determined with the YouTube data. The reason between the discrepancy between these two values of t may be explained by the contributions of the Truveo Search algorithm. As this data was collected from the Truveo Search engine, it is possible that the Truveo Search engine results generated by individuals using the service could affect the value of t. While further test methods and data analysis would be required to prove this, it is clear that the Zipf-like distribution still explains the behavior for the access frequency of

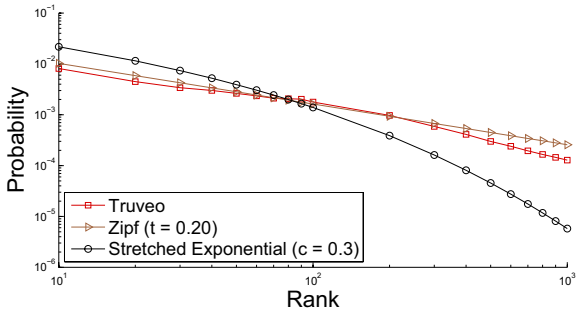
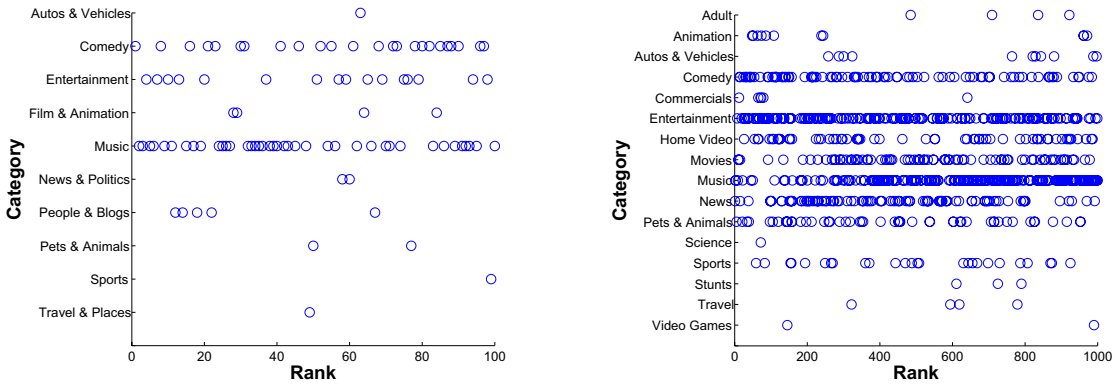


Figure 2. Truveo Video Search Probability Distribution

videos acquired through Truveo Search engine data. In contrast to study [9], Figure 2 illustrates that the data video access frequency does not follow the stretched exponential distribution.

4.2 Categories

The popularity of a video on social video websites may be caused by many factors. One of the first factors we examine is the type - or category - of the video that is published. Previous papers have showed that this video characteristic affects access frequency,[6,10] yet no paper thus far has surveyed all categories on a public social video website. We do so in this subsection and also analyze the relationship between video category and length. In Subsection 4.6, we compare these categories by region to study social aspects.



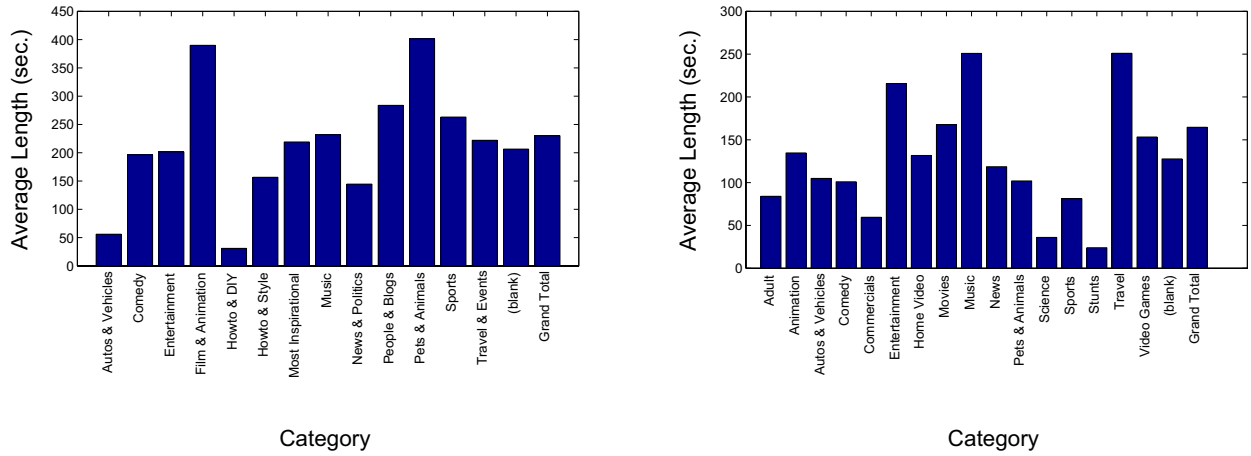
(a) YouTube (b) Truveo

Figure 3. Category vs. Rank Distribution

The video popularity distribution by category for the Top 100 YouTube Yearly Videos list is shown in Figure 3(a). Note that the majority of videos in this list are categorized as music videos, comedic videos, or entertainment videos. While a few comedic and entertainment videos have some of the top popularity, music videos are the most popular and have the most frequency of occurrence across the entire rank range on YouTube. Comedic videos have the second most frequent occurrence, yet most of the videos in this category have rank of 50 or lower. The third most frequent category, entertainment, has a similar trend. Autos & Vehicles, Travel & Places, and Sports each have the lowest access frequency, with Sports trailing all categories in terms of rank as well.

A similar trend can be observed for the Top 1000 Truveo Video Search list as shown in Figure 3(b). Comedy, entertainment and music videos again dominate the majority of videos published and accessed by users. Entertainment videos have the most frequency of occurrence across the entire rank range. Most comedic videos tend to have high ranking, whereas most music have low ranking. News is much more frequently accessed in Figure 3(b) in comparison to Figure 3(a) - most

notably in the 200 through 1000 rank range. As new categories are introduced with this dataset, Autos & Vehicles, Sports, and Travel & Places are no longer last in terms of categorical frequency - although their access remains sparse in the Truveo Search dataset. Science and Video Games are the two categories with the least number of videos in the Top 1000 list, even though both categories have a video with a rank better than 200.



(a) Top 100 YouTube Yearly Videos

(b) Top 1000 Truveo Videos

Figure 4. Average Video Length

It is interesting to note the relationship between the video length and category. As shown in Figure 4(a), Film & Animation, Pets & Animals, People & Blogs, and Sports categories all have video length averages over four minutes. As these video categories have very low frequency (see Figure 3(a)), it is evident that their larger video length averages do not have a significant impact on the overall average video length, compared with the Music category, which has an average of approximately 3 minutes and 45 seconds. Likewise, Figure 4(b) shows that entertainment and music videos in the Truveo Search data have an average length between 200 and 250 seconds. The overall average video length, however, is more dominated by the more popular comedic videos (approx. 100 and 115 seconds).

We analyze video length further later in this paper to identify more general trends across the rank distribution.

4.3 Days of the Week and Evolution

Let us now discuss how a video's popularity changes over time, as well as how the video access changes based on the day of the week and the time during the day. On August 23, 2007, we selected the Top 100 YouTube videos and froze the list for analysis. When analyzing the total daily view count for this list over an eight-week period, we observe interesting trends. First, Figure 5 shows that video popularity is nearly triple when analyzing accesses between 9pm - 1pm coordinated universal time (UTC) in comparison to the remaining twelve hours of the day. Further, Friday and Saturdays in general appear to be the most popular times for video access frequency, while Wednesdays have the least access frequency contribution. This is the first set of data known to show social video access frequency by day of the week and period of the day over such a long period.

Furthermore, our analysis shows that the access frequency of the Top 100 Videos decreases over time as illustrated in Figure 6(a). The total accesses for the Top 100 Videos decreased by nearly one million views per day from August 23rd to October 17th. The trends observed in the average view counts per day of the week are observed in similar fashion over the eight week period, where Fridays and Saturdays have the most views and dropping to a minimum for the week once Tuesday and Wednesday are encountered. This oscillating nature of video access frequency is very important to consider when planning for necessary bandwidth and load balance on social media websites such as YouTube.

Understanding how the Top 100 videos evolve based on rank provides answers regarding the video access frequency declines observed over the eight-week period. The rate of change in the accumulative view count is show in Figure 6(b).

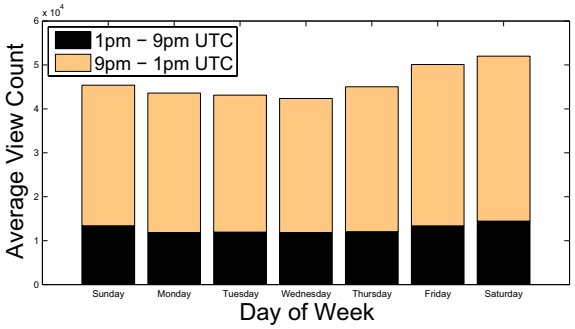
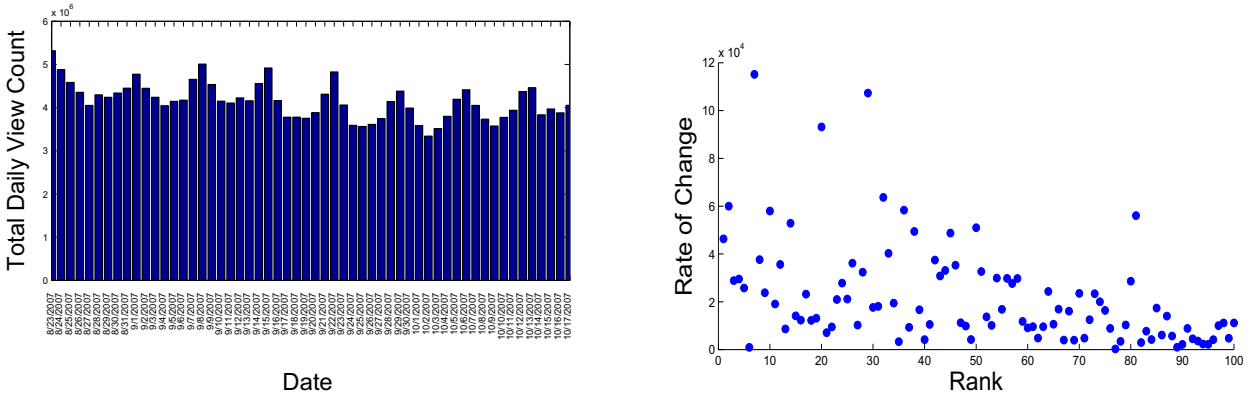


Figure 5. Total View Count Per Day of Week (Top 100 YouTube Videos, Averaged for Each Day over 8-Week Period)

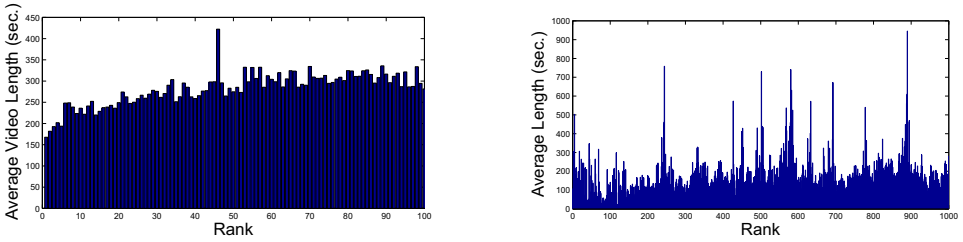


(a) Total Daily View Count (8-Week Period)

(b) Rate of Change in the Accumulative View Count

Figure 6. Top 100 YouTube Videos

Approximately 10% of the videos in the Top 50 continue to have frequent views by users; however, with a correlation coefficient of -0.465 (medium negative correlation), the general trend is less frequent access. In particular, the videos from rank 50 to 100 show a low rate of change across the period. It is further interesting to note that 6 out of the 10 videos with the greatest rate of change in Figure 6(b) are music videos. Popular music videos appear to maintain their access frequency over longer periods of time compared to other video categories.



(a) Top 100 Daily YouTube Videos

(b) Top 1000 Videos Indexed by Truveo

Figure 7. Average Length vs. Rank

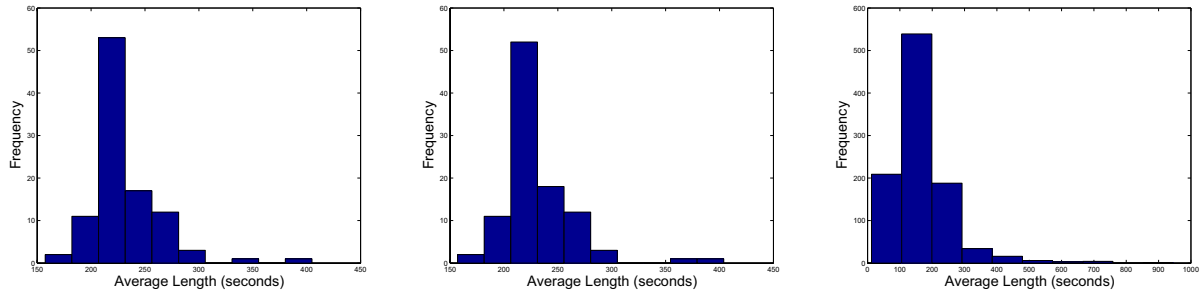
4.4 Length

The length of a video is very important to understand when planning client sessions with multimedia systems. As YouTube does not allow videos more than ten minutes in length, user sessions on YouTube are limited in time. This is much different

from the lengthy videos studied in [10]. Figure 7(a) shows the average video length profile for daily videos accessed. While rank does not appear to have a direct correlation to video length, it is interesting to note that 5 of the top 6 rankings have videos of two and a half minutes or less.

Videos greater than six minutes do not appear to be the norm in Figure 7(a), with the majority of videos having a length of five minutes or less. The correlation coefficient is 0.340 in Figure 7(a), which shows a tendency for videos of lower rank to be more lengthy in comparison to videos of higher rank. As you can see in Figures 8(a) and 8(b), a large portion of the most popular videos are in the three to six minute range.

More than 40% of the popular videos are concentrated in the four to five minute range in Figure 8(b). This is very important information, as it has direct implications on user access time, necessary bandwidth, and storage maintenance for the video's lifecycle. Videos are stored and available to be accessed on YouTube until the publisher removes them (or until a terms of use violation is realized). When analyzing the average video length for a larger set of videos, we observe a difference in the video population. Figure 7(b) shows that the Truveo Video Search engine indexes videos greater than ten minutes in length, different than the YouTube restriction. There are actually several videos - notably less than rank 200 - that are present in Figure 7(b). Second, a majority of videos across the distribution are below 250 seconds in length. Further, a majority of the videos with a rank of 100 or less appear to be less than 200 seconds in length in Figure 7(b). The correlation coefficient is 0.274, which shows small positive correlation and a slight tendency for videos of lower rank to be more lengthy than those of higher rank.



(a) Top 100 Monthly YouTube Videos (b) Top 100 Yearly YouTube Videos (c) Top 1000 Truveo Search Videos

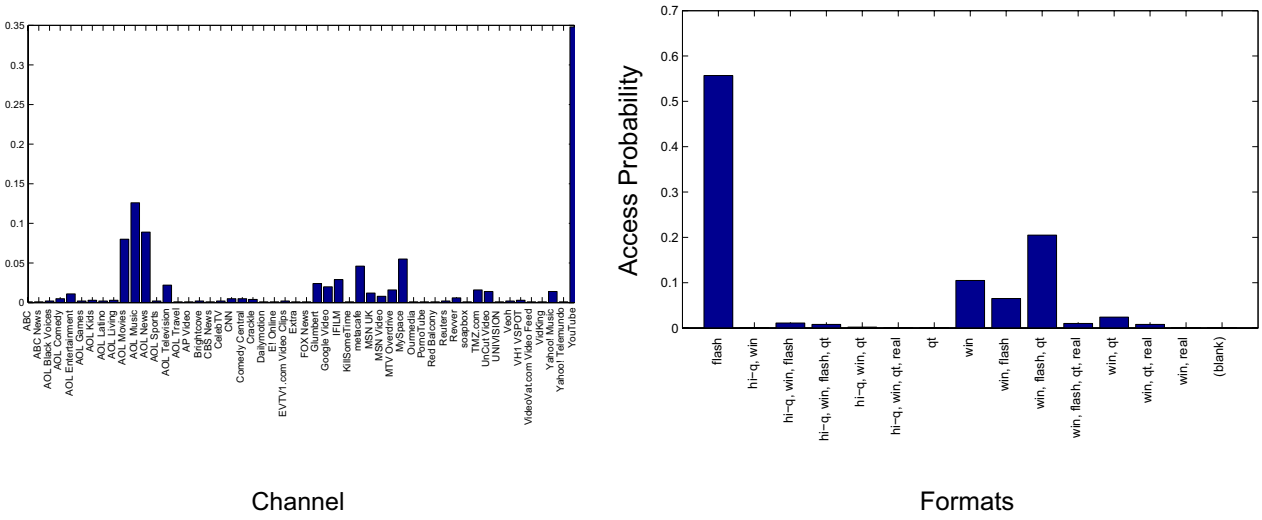
Figure 8. Average Length Distribution

Figure 8(c) shows a histogram of the Top 1000 videos from the Truveo Video Search engine. The frequency of videos in the 100 to 200 second range accounts for more than 50% of the videos over the distribution. There are several videos over five minutes in length; however, the frequency of this occurrence is much less in comparison to the norm.

4.5 Channels and Video Formats

We now further analyze two additional video characteristics: channel and video format frequency. As discussed earlier, YouTube is now the largest social video website and the third largest website accessed in the world. This is further confirmed by the channel distribution from the Truveo Video Search engine data. As we see in Figure 9(a), the YouTube channel (or, source of videos) accounts for 35% of video access alone, while the AOL Music, Movies, and News video channels together account for nearly 30% of the videos indexed by Truveo. MySpace, Metacafe, iFilm, Google Video, and Glumbert are some of the other channels that make up the top 90% of all videos accessed on the web. This channel distribution is significant, as it shows us that a large majority of users rely much more on YouTube than other websites for their online videos. The channel frequency distribution, therefore, from this data not only might show the potential necessity for resource allocation per channel, but also where opportunity for channel growth and consolidation of resources might lie.

Further analysis of the Truveo Video Search data shows a startling imbalance in video formats accessed by users (see Figure 9(b)). The Flash video format is the sole format for 55% of the videos published online. Further, more than 80% of videos published online can be provided to users in this format. Flash dominates Real Media, Windows Media, and QuickTime formats in Figure 9(b). This is significant data, as it allows for network engineers to anticipate necessary



(a) Channel Distribution

(b) Video Format Distribution

Figure 9. Top 1000 Videos Indexed by Truveo

bandwidth loads, for consumers of the content to ensure they have the latest updated codecs and players, and ultimately content providers to move toward the most relevant media type to maintain and attract consumers.

4.6 Geographical Aspects

Video popularity and rank up until this point have been aggregated and presented without specific details regarding the users that are participating in social video websites. Understanding the background of people who participate in social video websites adds an entirely new dimension to understanding a particular culture. The analysis of Google Video data provides this added dimension and helps to understand which types of videos users from different regions around the world like to watch.

The distribution of video popularity for the Top 100 Google Videos for all countries and regions is shown in Figure 10. This Google Video data confirms again that music, comedic, and entertainment videos are the most popular. Here, music is nearly twice as popular as the comedy and entertainment categories. Different from our previous YouTube category analysis in this paper, sports, short movies, blogs, and TV shows are the next most significant categories that are watched, each accounting for at least 5% of the Top 100 videos.

As the data from Google Video is based on videos from Google, YouTube, and a few other video websites, we get a significant picture of what people around the world like to watch in Figure 10. The data in Figure 10 is biased, however, by the total number of view counts and does not give any information on what people in the region of Oceania like to watch in comparison to people in the Near East and how video popularity among categories might differ based on geographical region. In order to get a much more detailed analysis, we analyzed data for the forty-four countries for which Google reports and grouped the data into the regions identified by the U.S. Census Bureau: Asia (excluding Near East), Commonwealth of Independent States, Eastern Europe, Latin America and the Carribean, Near East, Northern America, Oceania, Sub-Saharan Africa, and Western Europe. We weighted the data collected for each region by the country from which it was attributed. We then further weighted the ranking by the Zipf distribution to ensure each category achieved appropriate importance in the analysis.

As can be seen in Table 1, music, comedy, and entertainment have video categories which are watched in all regions. Their favor in each region varies considerably, however. Music in Latin America and the Carribean, Northern America, and Western Europe each account for more than 30% of the videos watched. This is in comparison with more than 20% influence in Sub-Saharan Africa and approximately 15% influence in the Commonwealth of Independent States and Eastern Europe, respectively. The only other video category that comes close to such an influence within a particular region is comedy, which accounts for 30% of the videos watched in Eastern Europe and 28% in Oceania, with further

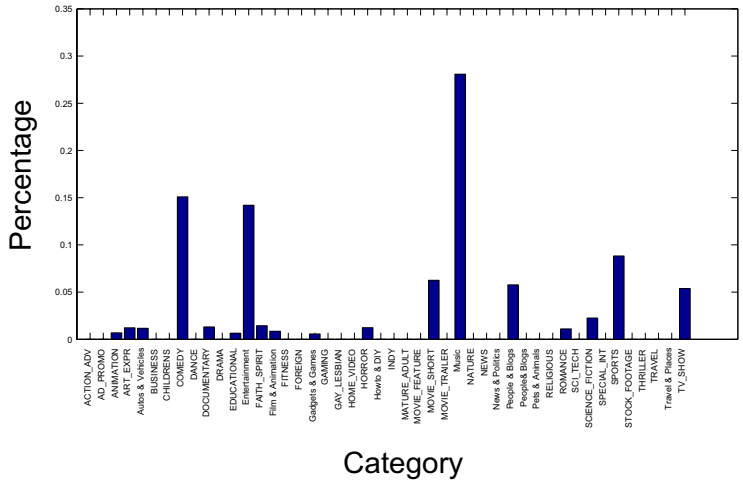


Figure 10. Video Category Distribution for Top 100 Google Videos

considerable influence in nearly all other regions. Entertainment videos have a wide array of influence. Most notably, 18% of videos watched in Asia (excluding Near East) are entertainment videos, while nearly 15% in Northern America and 13% in Sub-Saharan Africa respectively.

Some very unique video interests can be observed in Table 1. First, we note that although music, comedy, and entertainment are most watched in Asia, short movies and sports videos also have considerable viewers in the region. Viewers in the Commonwealth of Independent States have considerable viewing interest in documentaries - nearly as much as comedy and music - and watch gaming and people and blogs videos more than any other region. Next, viewers in the Near East view feature movies most frequently - more than eight times any other region - and have considerable interest in sports, action adventure, and short movies. While Northern America mostly watches music videos, Oceania is most influenced by comedic videos and has an opposite priority of viewing entertainment and music videos accordingly. Lastly, Sub-Saharan Africa watches sports videos more than any other region and this video category influence is only second in the region to music.

Further study around the reasons for these video choices need to be determined; however, we present these findings as a starting point for further analysis and study. Cultural study and sociological reasoning may indeed be necessary to analyze the importance of social video genres and their content. What we do present is the characterization of these video categories on popular videos around the world and emphasize the distribution of video served to those regions.

5. CONCLUSION

This paper has investigated online social video content from a global perspective. The results from over seven months of data collection demonstrate that video popularity on YouTube follows Zipf-like distribution with $t = 0.30$ for daily access and $t = 0.50$ for weekly, monthly, and yearly access, while video popularity from the Truveo Video Search data follows a Zipf-like distribution with $t = 0.30$. In the category analysis, we have demonstrated that music, comedic, and entertainment videos are the most popular. When correlating both category and video length, we have shown that the average video length of these popular videos have a significant effect on the total average video length of the YouTube and Truveo videos. Access patterns demonstrate an oscillating curve as video popularity varies per day of the week, while the most frequent views happen on Fridays and Saturdays during 9PM and 1PM UTC. The rate of change in the access of the most popular videos exhibits medium negative correlation with the rank. Finally, we have provided extensive popularity distributions based on upon the nine regions of the world defined by the U.S. Census Bureau. Using Google Video data, we have provided comparative data on the video interests of those nine regions.

		REGION								
		Asia (excl. Near East)	Common. of Indep. States	Eastern Europe	Latin America & Carribean	Near East	Northern America	Oceania	Sub-Saharan Africa	Western Europe
CATEGORY	Action & Adventure	0.30%	3.20%	1.02%	0.05%	8.21%	0.00%	0.38%	1.06%	0.23%
	Anime & Animation	1.54%	0.00%	7.21%	1.63%	0.82%	0.66%	1.91%	0.72%	3.57%
	Art & Experimental	0.90%	1.15%	0.61%	2.22%	0.04%	1.59%	0.24%	2.04%	2.71%
	Autos & Vehicles	0.48%	1.32%	0.67%	1.15%	0.47%	0.89%	0.12%	2.98%	1.13%
	Children & Family	0.16%	0.83%	1.39%	0.18%	0.05%	0.00%	0.00%	0.00%	0.16%
	Comedy	9.24%	15.58%	30.04%	12.15%	13.83%	15.09%	28.14%	14.49%	13.89%
	Documentary	1.74%	13.73%	1.55%	1.50%	1.80%	0.87%	2.60%	1.17%	1.88%
	Drama	0.40%	1.57%	0.00%	0.18%	2.40%	0.00%	0.00%	0.00%	0.20%
	Educational	1.62%	0.53%	0.00%	1.25%	2.15%	0.66%	1.64%	0.00%	1.17%
	Entertainment	18.16%	5.29%	1.53%	8.63%	8.03%	14.64%	10.93%	12.63%	8.12%
	Faith & Spirituality	0.54%	0.44%	0.63%	0.21%	0.06%	2.03%	3.14%	1.51%	0.52%
	Film & Animation	2.17%	0.37%	2.54%	1.82%	0.51%	1.17%	0.62%	0.00%	0.57%
	Foreign	0.71%	1.39%	0.00%	0.03%	0.19%	0.00%	0.46%	0.00%	0.13%
	Gadgets & Games	0.40%	1.31%	0.00%	0.08%	0.00%	0.84%	0.60%	0.00%	0.05%
	Gaming	3.04%	7.00%	0.58%	2.24%	0.37%	0.18%	0.14%	0.75%	1.86%
	Home Video	0.49%	1.09%	2.62%	0.37%	3.51%	0.65%	0.37%	1.00%	0.64%
	Horror	0.86%	0.00%	0.75%	0.08%	0.00%	1.22%	2.41%	0.45%	0.33%
	Howto & DIY	0.61%	0.69%	0.00%	0.76%	0.06%	0.00%	0.11%	1.97%	0.26%
	Movie (feature)	2.61%	1.86%	1.05%	1.62%	16.86%	0.06%	0.69%	0.91%	1.76%
	Movie (short)	7.83%	4.42%	11.08%	4.38%	9.11%	4.29%	6.38%	3.24%	4.87%
	Movie Trailer	0.60%	0.00%	0.00%	0.27%	2.87%	0.00%	0.19%	0.63%	0.57%
	Music & Musical	12.10%	15.25%	15.45%	34.17%	3.91%	31.90%	8.43%	22.24%	31.90%
	News	0.47%	0.00%	0.00%	1.39%	0.02%	0.55%	0.00%	0.60%	0.03%
	News & Politics	2.88%	1.46%	0.00%	0.68%	0.45%	0.67%	0.00%	0.00%	0.11%
	People & Blogs	5.32%	7.68%	2.28%	3.86%	4.19%	7.21%	3.56%	1.78%	3.81%
	Pets & Animals	0.06%	0.00%	0.00%	0.00%	0.21%	0.13%	2.67%	1.70%	0.42%
	Romance	4.07%	0.87%	0.00%	0.35%	1.36%	0.95%	0.68%	0.53%	0.23%
	Science & Technology	0.50%	0.42%	0.00%	0.11%	0.00%	0.06%	2.16%	0.35%	0.32%
	Sci-Fi & Fantasy	2.87%	1.18%	0.00%	0.30%	0.27%	1.59%	3.20%	0.00%	0.60%
	Special Interest	0.32%	1.82%	0.00%	0.60%	0.35%	0.00%	0.56%	0.00%	0.51%
	Sports	7.40%	4.35%	9.70%	8.50%	10.76%	2.34%	6.79%	18.31%	10.06%
	TV Show	3.83%	4.01%	5.52%	4.69%	3.22%	5.23%	6.33%	3.88%	3.50%

Table 1. Regional Video Category Comparison

REFERENCES

[1] "Flickr growth and bumped images." http://www.flickr.com/photos/gustavog/2324101087/.

[2] "Online gaming revenues to triple by 2009." `http://www.parksassociates.com/press/press_releases/2005/gaming-1.html`.

[3] "Breakthrough internet device." `http://www.apple.com/iphone/features/index.html#internet`.

[4] "Alexa, the web information company." `http://www.alexa.com`.

[5] Zink, M., Suh, K., Gu, Y., and Kurose, J., "Watch global, cache local: Youtube network traffic at a campus network measurements and implications," in [*Multimedia Computing and Networking (MMCN)*], (January 2008).

[6] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., and Moon, S., "I tube, you tube, everybody tubes: analyzing the worlds largest user generated content video system," in [*ACM Internet measurement Conference (IMC)*], (October 2007).

[7] Gill, P., Arlitt, M., Li, Z., and Mahanti, A., "Characterizing user sessions on youtube," in [*Multimedia Computing and Networking (MMCN)*], (January 2008).

[8] Duarte, F., Benevenuto, F., Almeida, V., and Almeida, J., "Geographical characterization of youtube: a latin american view," in [*LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*], 13–21, IEEE Computer Society, Washington, DC, USA (2007).

[9] Guo, L., Tan, E., Chen, S., Xiao, Z., and Zhang, X., "Does internet media traffic really follow zipf-like distribution?," *SIGMETRICS Perform. Eval. Rev.* **35**(1), 359–360 (2007).

[10] Acharya, S., Smith, B., and Parns, P., "Characterizing user access to videos on the world wide web," in [*Multimedia Computing and Networking Conf. (MMCN)*], (January 2000).

[11] Chesire, M., Wolman, A., Voelker, G., and Levy, H., "Measurement and analysis of a streaming media workload," in [*USENIX Symposium on Internet Technologies and Systems (USITS)*], 1–12 (March 2001).

[12] Almeida, J. M., Krueger, J., Eager, D., and Vernon, M., "Analysis of educational media server workloads," in [*11th international workshop on network and operating systems support for digital audio and video (NOSSDAV)*], 21–30 (June 2001).

[13] Cherkasova, L. and Gupta, M., "Characterizing locality, evolution, and life span of accesses in enterprise media server workloads," in [*12th Int'l. Workshop on Network and Operating System Support for Digital Audio and Video (ACM NOSSDAV 2002)*], (2002).

[14] "Truveo video search." `http://developer.truveo.com/index.php`.

[15] "Google-video project." `http://rubyforge.org/projects/google-video/`.

[16] "Youtube project." `http://rubyforge.org/projects/youtube`.

[17] "U.s. census bureau international data base." `http://www.census.gov/ipc/www/idb/tables.html`.