Understanding Near-Duplicate Videos: A User-Centric Approach

Mauro Cherubini, Rodrigo de Oliveira, and Nuria Oliver
Telefónica Research
via Augusta, 177 – 08021 Barcelona, Spain
{mauro, oliveira, nuriao}@tid.es

ABSTRACT

Popular content in video sharing web sites (*e.g.*, YouTube) is usually duplicated. Most scholars define near-duplicate video clips (NDVC) based on non-semantic features (*e.g.*, different image/audio quality), while a few also include semantic features (different videos of similar content). However, it is unclear what features contribute to the human perception of similar videos. Findings of two large scale online surveys (N = 1003) confirm the relevance of both types of features. While some of our findings confirm the adopted definitions of NDVC, other findings are surprising. For example, videos that vary in visual content –by overlaying or inserting additional information– may not be perceived as near-duplicate versions of the original videos. Conversely, two different videos with distinct sounds, people, and scenarios were considered to be NDVC because they shared the same semantics (none of the pairs had additional information). Furthermore, the exact role played by semantics in relation to the features that make videos alike is still an open question. In most cases, participants preferred to see only one of the NDVC in the search results of a video search query and they were more tolerant to changes in the audio than in the video tracks. Finally, we propose a user-centric NDVC definition and present implications for how duplicate content should be dealt with by video sharing websites.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Search process*

General Terms: Human Factors

Keywords: NDVC, psychophysical experiment, similarity, user study, video sharing, YouTube

1. INTRODUCTION & MOTIVATION

Today's video-sharing web sites allow their users to freely post multimedia content without typically checking for uniqueness. As a consequence, it is not unusual to find in these sites multiple copies of the same or very similar videos. These videos are usually referred to as *near-duplicate video clips* (NDVC).

Different research groups have related the presence of NDVC with spam creation [2] and copyright infringements [14]. For example, Wu *et al.* [16] recommend the identification and removal of this duplicated content in order to increase the efficiency of video information retrieval tasks. In their studies, they found an average of 27% of NDVC in the search results of an original video.

Most of the previous work in this area has focused on identifying and removing NDVC. However, we believe that these approaches understate the role played by NDVC, as they are not necessarily uploaded with malicious intent or are exact copies of the original video. In fact, it is not infrequent to find near-duplicate clips that *complement* the original material with additional information (*e.g.*, commentary audio or subtitles) that might provide valuable information to the users of the system. Furthermore, there does not seem to be a full agreement on the technical definition of the features that characterize NDVC.

Therefore, we believe that the multimedia information retrieval community would benefit from additional human-centric research on this topic – gathered via user studies, for at least three reasons: 1) Little is known about how users are affected by the presence of NDVC; 2) it is generally unknown what features contribute to the users' perception of similarity among multimedia items; and 3) there is a lack of empirical proofs showing that the removal of NDVC from the results set of a video search task satisfies the users' needs.

In this paper, we present the results of two large-scale online questionnaires that were designed to shed light on the human perception of NDVC. We asked respondents to:

1. characterize their common use of video sharing websites;
2. watch pairs of NDVC and state their degree of similarity (pairs differed in only one feature);
3. elicit their preferences –if any– on which duplicate they would like to have in the search results (see Sections 3 and 4).

These measurements led us to a user-centric definition of NDVC with implications for how duplicated videos should be retrieved in video-sharing web sites (see Section 5).

2. RELATED WORK

In the last few years, different research groups have tried to understand how video sharing web sites are used. A large part of the work has focused on YouTube[1], the largest and most popular video sharing website today. The focus has been on gathering objective measurements of the users' interactions in these sites, mainly with two goals in

[1]See `http://www.youtube.com`, lastly retrieved in April 2009.

MM'09, October 19–24, 2009, Beijing, China.

Telefónica I+D participates in Torres Quevedo subprogram (MICINN), cofinanced by the European Social Fund, for Researchers recruitment.

mind: 1) improving the efficacy of the video information retrieval task; and 2) fighting malicious behavior such as spam, self-promotion of certain users, and copyright infringements. First, we shall review the most relevant work that analyzes the behavior of users of video sharing sites (particularly YouTube), followed by an overview of the literature in NDVC detection and removal.

2.1 Analyzing YouTube User Behavior

Benevenuto *et al.* [2] conducted a study to understand user behavior on YouTube. In particular, they crawled YouTube and studied how people interact with each other through video responses by measuring degree distributions in their interactions. They found that 60% of YouTube users have an out-degree higher than in-degree, whereas only 5% of the users have significantly higher in-degree than out-degree. In other words, a very small number of users act as authorities or *hubs* of information (those with high in-degree) while the majority of users are low-rank users, have a small number of views and receive none or very few video responses from the video community. Using the same approach they found consistent evidences of anti-social behavior. For example, nodes with very high out-degree may indicate either very active users or spammers.

Complementary results were obtained by Halvey and Keane [7] who analyzed social interactions on YouTube by crawling user pages and focusing on website-supported methods for social interactions. They found that users tend to watch rather than to add videos (*e.g.,* 966 views *vs.* 11 uploads on average per user). Furthermore, they found a general failure in exploiting the community facilities available on the website. These findings are very relevant when designing a personalization or recommendation system for YouTube users, as this passive user behavior might not be informative enough for generating predictions for a community of users. Similar results were obtained by Gill *et al.* [5] who, following a similar methodology, found that most users do not upload videos (*e.g.,* 51% of sessions did not transfer any videos) and have different browsing patterns depending on the purpose of their visit. Finally, a finer profiling of YouTube users was described by Maia *et al.* [10] where they collected a large dataset containing many features of the users' interactions in the system. They then clustered users into 5 user types. Out of their sample, only 23% of the users were identified as *content producers*, *i.e.*, users that constantly access their accounts and have a significantly higher than average number of uploads, watches, and channel views.

A common pattern found in these studies is that the *greatest part of the users of video sharing web sites consume media instead of sharing it.* However, little research has been carried out to date on *how* users reach the content they watch. This specific point is relevant to understand what population of users is affected by the problem of near-duplicate videos. Note that users who access videos by following recommended links will not experience the presence of near duplicates. Conversely, users who actively search for video content will be exposed to NDVC in their search results. Hence, we formulate our first hypothesis as: *H1, Video search is the main method for reaching content on video sharing websites.*

2.2 Near-Duplicate Video Clips

Turning now our attention to NDVC, it is important to understand the role that duplicated clips play on the way users use video sharing web sites. In this regard, Kruitbosch and Nack [9] investigated to what extent the videos shared on YouTube are self/amateur generated content *vs.* professionally authored content. They found that most of the *popular* content on YouTube was professionally generated, even though a random sample showed that there was significantly more user-generated content available. In this sense, YouTube seems to be acting as a social filter, allowing anyone to share content they find interesting, rather than a way for creative people to show their abilities to the world. Professionally created videos are more likely to be copied than user-generated ones [9]. Given that most of the popular content in video sharing web sites has been found to be professionally generated, one would expect to find a significant number of NDVC in these sites.

Cha *et al.* [3] conducted several experiments on a large dataset of YouTube videos. They found that the way content is filtered on YouTube is the likely cause for the lower-than-expected popularity of niche contents, which if leveraged could increase the total views by as much as 45%. More specifically, they conducted experiments to understand the impact of content aliasing. They extracted a sample of 216 of the top 10.000 videos on YouTube and found that about 85% of them had 1 to 4 duplicates. Most of the duplicated videos were uploaded on the same day as the original video or within a week. In addition, many of them still appeared 100 or more days after the original videos were posted. Less dramatic results were reported by Wu *et. al.* [16] who conducted a study on the topmost search results on a sample of 24 popular queries from YouTube, Google Video and Yahoo! Video. They found an average of 27% NDVC of the most popular version of a video in the search results. These results suggest that the presence of NDVC in the search results is a real problem that impacts the way people reach for content on video sharing websites. Note that in all studies NDVC are seen as *redundant* content and therefore have attracted the attention of multimedia information retrieval scholars who have proposed in recent years approaches to detect and cluster or eliminate NDVC from the search results.

The first step when building a NDVC detection system is a working definition of NDVC. Table 1 summarizes the most common definitions of NDVC that have been proposed in the literature. As seen on the Table, the actual definition of NDVC is still an open research question. We summarize next the most relevant and recent efforts –and associated NDVC definitions– in automatically detecting NDVC from a video search result list.

Wu *et al.* [16] tried to identify and remove NDVC using the definition reported in Table 1. They proposed a hierarchical approach to cluster and filter out NDVC, demonstrating that their approach could effectively detect and reduce redundant videos displayed to the user in the top result set. Shen *et al.* [14] extended the definition of NDVC by including changes introduced during capturing time, such as a change of camera view point (see Table 1). They proposed a detection system called UQLIPS that comprised two approaches: a bounded coordinate system and a frame symbolization, which takes temporal order of the key-frames into consideration. They found that this system could accurately remove NDVC from a large collection in real-time. Yet another definition was employed by Basharat *et al.* [1], who included intra-class variations such as scene settings, different viewpoints, different camera motions, to name a few.

Taking as a starting point all previous work, we devised an experiment to test –from a user-centric perspective– which of the features proposed in the literature play a role in the users' perceptions of NDVC. Therefore, we pose our second hypothesis as: *H2, Identical or approximately identical videos differing in photometric features (image quality), audio quality, editing of the content (i.e., few or more scenes), additional content (i.e., audio and image overlays), or hav-*

ing the same visual context but different audio (or viceversa) are considered by the users as similar clips. Finally, we seek to verify our initial argument that users might not want to have all this duplicated content removed from the search results. Hence, our third hypothesis is: *H3, Once the users obtain the result list fo a video search query and after watching the NDVC in such a list, they have a preference for one NDVC over the others and therefore would rather only see the preferred NDVC in the results.*

Table 1: Comparison of NDVC definitions

Author	NDVC definition
Wu et al. [16]	Identical or approximately identical videos close to the exact duplicate of each other, but different in file formats, encoding parameters, photometric variations (color, lighting changes), editing operations (caption, logo and border insetion), different lenghts, and certain modifications (frames add/remove).
Shen et al. [14]	Clips that are similar or nearly duplicate of each other, but appear differently due to various changes introduced during capturing time (camera view point and setting, lighting condition, background, foreground, *etc.*), transformations (video format, frame rate, resize, shift, crop, gamma, contrast, brightness, saturation, blur, age, sharpen, *etc.*), and editing operations (frame insertion, deletion, swap and content modification).
Basharat et al. [1]	Videos of the same scene (*e.g.*, a person riding a bike) varying viewpoints, sizes, appearances, bicycle type, and camera motions. The same semantic concept can occur under different illumination, appearance, and scene settings, just to name a few.

We believe that the previous work in this area has been extremely valuable, but would greatly benefit from a user study focused on the needs and perceptions of users of video sharing sites.

An underlying challenge in this research is related to the subjectivity of the human perception [12, 13]: different users might have different reactions to a particular definition of NDVC and might have different preferences on how to treat this content (*e.g.*, hide it *vs.* cluster it). In the field of image retrieval, recent psychophysical experiments have been conducted to capture the users' perceptions and to use them as the ground truth when evaluating the performance of retrieval algorithms. In all the studies, the retrieval performance was significantly improved by incorporating the human perception of similarity into the systems [11, 6, 4], thus highlighting the importance of extending user studies of human perception to video similarity as well.

In summary, the work presented in this paper aims at providing evidences on: 1) how users of video-sharing web sites reach the content they intend to watch; 2) whether different features that are used to characterize NDVC are perceived as potentially producing redundant content; and 3) whether users have preferences on the way they treat NDVC.

We believe that finding the answers to these three points will be instrumental in the development of efficient, useful and intuitive search and retrieval systems of audiovisual content.

3. METHODOLOGY

The ultimate user of a video retrieval system is a human being. Therefore, the study of the perception of video content from a psychophysical perspective is of crucial importance. In our work, we have conducted a psychophysical experiment to measure the perceived similarity of NDVC by collecting a large number of subjective answers on video similarity. We presented pairs of videos to subjects using a technique similar to that used in the past for measuring image similarity [11, 6, 4]. We wanted the experiment to take place in an ecologically valid environment. Thus, we opted for an online questionnaire technique instead of an in-lab study. Note that streamed videos are usually watched in displays with different sizes, resolutions, and contrast levels. Therefore, the online setting would allow participants to compare videos using their usual configuration.

The study was designed to test each of the three hypothesis presented in Section 2. In terms of **H1**, we investigated the users' behavior in a video search task from two perspectives: *purpose* and *proactivity*. With respect to purpose, subjects were asked to report the most common tasks that they performed in video sharing websites such as YouTube: (1) search for specific videos, (2) browse without a specific video in mind, or (3) do something else. In terms of proactivity, participants answered if the videos they watch on these systems are usually: (1) found by themselves, (2) suggested by someone else, or (3) found by other means.

Concerning **H2**, we asked participants to watch seven pairs of NDVC, where each pair of NDVC differed in only one feature, as detailed in Section 3.2. Subjects were asked to rate the similarity of the paired videos and to state *why* they chose a particular degree of similarity.

Finally, **H3** was addressed by asking participants: 1) whether they had a preference between each of the paired videos; and 2) which of the videos they would like to see in the result set if they were searching for videos using the same query. Answers were limited to: (1) video 1, (2) video 2, (3) both, (4) none, (5) either one, and (6) "I don't know what to expect from the query associated with the videos".

3.1 Procedure

To test our hypothesis, we deployed a large-scale questionnaire on one of the most visited news portals in Spain[2]. Visitors of the portal could see a banner on the front page that advertised our research initiative. After clicking on the banner, they were redirected to the online questionnaire. The system that hosted the form registered the IP of the respondents and the timestamps at which each respondent started and ended answering the questions. As an incentive, three 100 euro vouchers were raffled among all respondents.

We deployed the questionnaire *twice* in order to collect both qualitative and quantitative information from participants while avoiding potential biases in the answers. The first deployment (Q1) lasted one week and the questions related to H1 and to the *why*-component of H2 were left as open questions. These qualitative answers were manually categorized at the end of the week and used to define multiple-choice questions in the second deployment of the questionnaire (Q2), which was available for two weeks. For example in Q1, after the participants defined the similarity between the clips of condition D we asked them to elaborate. A typical answer was: "they are different because one has a commentary and the other does not". In Q2, this was translated in the choice: "I noted relevant differences between the videos".

In order to validate H2, we selected NDVC examples from YouTube following the procedure described in Section 3.2. The presentation order of the video examples followed a Latin square design to avoid bias, thus creating seven groups (*i.e.*, ABCDEFG, GABCDEF, FGABCDE, and so forth).

[2]See `http://www.terra.es`, lastly retrieved in April 2009.

Each participant was submitted randomly to only one group.

For each of the seven pairs (conditions), participants were required to fully watch both videos at least once, and rate how similar they thought these videos were using a 5-point Likert scale. Participants could watch the videos as many times as they liked. All videos had an associated audio track.

3.2 Stimuli

In order to validate H2, we selected the most viewed videos on YouTube from "last month" and "at all times", excluded those with inappropriate content (*e.g.*, accidents, pornography, *etc.*), and created queries to retrieve the remaining videos.

From the results set, we identified five NDVC pairs that exemplified variations of the most common non-semantic features [14, 16], and two pairs that illustrated variations of semantic features [1]. The selected videos were edited such that all NDVC pairs would have about the same length ($\overline{x} = 37$ seconds), except in condition C (see Table 2)[3].

3.3 Participants

A pool of 647 participants answered Q1 while 553 answered Q2. In terms of validating H1, we considered only subjects that answered all questions related to how they use video sharing websites. A total of 498 subjects (270 male, 228 female) from Q1 and 505 subjects (286 male, 219 female) from Q2 complied with this requirement. The median age was 30 years (min: 12, max: 81) and 32 years (min: 12, max: 63) for Q1 and Q2 respectively. Both samples had more than 97% of Spanish subjects with a wide range of occupations. Subjects reported using computers everyday and samples seemed to differ regarding how frequently they use video sharing web sites (from 4 to 6 days a week in Q1, and everyday in Q2). However, this difference was not significant ($p = .34$). With respect to the most used system, YouTube was clearly the most popular (98% in Q1 and 96% in Q2), followed by Google Video (54% in Q1 and 65% in Q2), MSN Video/Soapbox (34% in Q1 and 32% in Q2), MySpace Video (33% in Q1 and 32% in Q2), and Yahoo! Video (26% in Q1 and 34% in Q2). While Terra TV was used by 66% of the Q2 sample, we decided not to include it in the analysis given that the questionnaires were deployed in this news portal, which probably biased the answers to this question.

In terms of validating H2 and H3, only 217 subjects from Q1 (105 male, 112 female) and 231 from Q2 (136 male, 95 female) complied with all of the requirements of the study: 1) fluent in Spanish; 2) experience with at least one video sharing website; 3) could listen to the audio track in the videos by means of the computer speakers or a headphone; 4) had no significant audio or video impairment; 5) took at least the minimum amount of time possible to fill out the questionnaire[4]; and 6) their answers had no missing data. Note that the Q1 and Q2 samples preserved the same amount of subjects per group regarding the presentation order of the video examples (see Section 3.1). Subjects' median age was 31 years (min: 16, max: 63) and 32 years (min: 18, max: 61) in Q1 and Q2 respectively. Both samples had more than 92% of Spanish subjects with a wide range of occupations, and reported using computers and video sharing web sites everyday.

[3]The clips used in this experiments can be viewed at: `http://tinyurl.com/youtubestudy`, last retrieved April 2009.

[4]Subjects took medians of 18 and 19 minutes to answer each questionnaire (Q1 and Q2 respectively). As 8.7 minutes were required to watch the 14 videos (7 NDVC pairs), we stipulated 10 minutes as the minimum.

3.4 Measures

Multiple choice questions with single answer were used to test both H1 and H3, whereas H2 was tested by means of 5-point Likert scale questions, designed to rate the similarity between the seven NDVC pairs. The textual explanations that the participants gave to each of their ratings in Q1 were manually categorized.

3.5 Statistical Analysis

In both Q1 and Q2, subjects were randomly distributed in seven groups in order to balance the presentation order of the seven NDVC pairs in a Latin square basis (see Section 3.1).

With respect to the validation of H1, we considered only subjects who answered questions related to *how* they use video sharing websites (Q1: $n = 498$, Q2: $n = 505$). With respect to the validation of H2 and H3, we used data from subjects that complied with all requirements presented in Section 3.3 (Q1: $n = 217$, Q2: $n = 231$). As the observed variables in both questionnaires were neither nominal or ordinal, we used the following non-parametric tests and measures to investigate the differences and possible associations between them:

– *Mann-Whitney U test (M-W test):* Used to assess whether two independent samples of observations at the ordinal level come from the same distribution (*e.g.*, the similarity level obtained for NDVC from condition A in Q1 constitutes one sample, while the same observation in Q2 constitutes the other sample);

– *Kolmogorov-Smirnov test (K-S test):* Used to complement findings obtained with the M-W test. The K-S test can also assess whether two independent samples of observations at the ordinal level share a number of properties between distributions [15] (*e.g.,* shape). As it makes no assumption about the distribution of the data, it is less likely to detect small differences in the median than the M-W test;

– *Chi-Square test (χ^2):* Used to assess whether two independent samples of observations at the nominal level come from the same distribution (*e.g.*, the user's preference over NDVC from condition A in Q1 constitutes one sample, while the same observation from Q2 constitutes the other sample). Other statistics derived from the Pearson Chi-Square were also used, such as the *Phi coefficient* (ϕ) to measure the association between two dichotomies (*e.g.*, v1: *find-video* –do participants watch videos found by themselves or suggested by someone else; v2: *have-account* –do users have an account on a video sharing website), and the *Contingency Coefficient* (C) to measure the association between two nominal/ordinal variables (*e.g.*, v1: *find-video* –do participants watch videos found by themselves or suggested by someone else; v2: *video-freq* –how frequently subjects use video sharing websites);

– *Spearman's Rho (ρ):* Used to measure the correlation between two ordinal related variables (*e.g.*, v1: similarity level between two NDVC from condition A; v2: subject's image expertise);

4. RESULTS AND DISCUSSION

4.1 Validation of H1

> *Video search is the main method for reaching content on video sharing websites.*

The following methodology was used in order to falsify this hypothesis. We started with the initial pool of respondents for each questionnaire and identified the number of subjects that use video sharing web sites (q1 below). For this set of respondents, we identified how many use these systems

Table 2: Descriptions of the seven NDVC pairs used in the questionnaires

Condition	Query	Video 1		Video 2	
A Photometric variation	crazy frog chanpions	A1: standard image		A2: higher quality (better colorfulness and lighting)	
B Editing operation (add/remove scenes)	skate Rodney Mullen	B1: fewer scenes, more content per scene		B2: more scenes, fewer content per scene	
C Different length	how to search in Google Maps	C1: first 38 seconds of video C2		C2: C1 with 24 seconds of extra content	
D Editing operation (audio, image overlays)	plane airport Bilbao wind	D1: no overlays		D2: overlays (audio comments and logo)	
E Audio quality	More than Words	E1: stereo audio in 44KHz		E2: mono audio in 11KHz	
F Similar images and different audio	atmospheric pressure	F1: experiment with a soda can		F2: experiment with a beer can	
G Similar audio and different images	Beatles all you need is love	G1: original musical clip		G2: G1 song performed by another band	

proactively (q2 below): *i.e.,* when they watch a video, it is usually a video that they found by themselves instead of being suggested by someone else. Finally, we highlight the fraction of these participants that usually have a *purpose* when searching for specific videos instead of browsing with nothing in mind (q3 below). If the proportion of *proactive* users is smaller than that of passive users, or if they do not search for videos more than they do any other task on video sharing websites, we reject the hypothesis. In addition, we present a tree graph that characterizes the profiles of user behavior in these systems (see Figure 1). The figure shows that most of the subjects access the videos in the video-sharing website by using keywords (thicker edges) rather than browsing categories on the main page.

q1. *"How many subjects use video sharing websites?"* From the initial pool of 634 respondents of Q1, 14 reported that they never used any video sharing website (2% of the participants). This figure was even smaller in Q2 (6 out of the 553 initial respondents). Both questionnaires were advertised to a target audience of users that watch videos on the web. Therefore, we assume that 98% constitutes the fraction of those who watch videos on the web that do it by means of video sharing websites. Hence, this majority of users will be prone to experience NDVC. Next, we identify how proactive are the users of these sites.

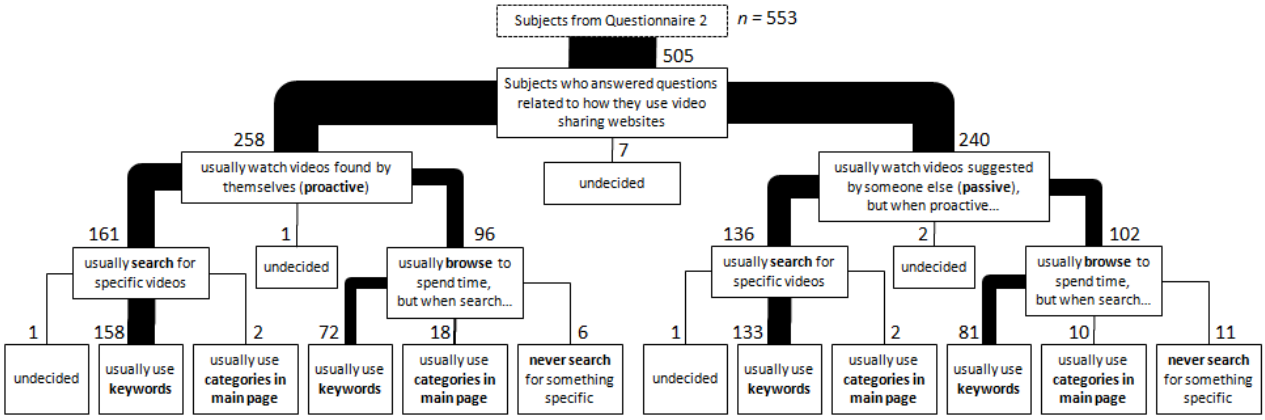
q2. *"How many subjects use video sharing websites proactively?"* With respect to Q1, from the 620 users of video sharing websites, 498 answered the questions related to their behavior when using these systems: 266 subjects (53%) reported watching videos found by themselves; 216 (or 43%) reported watching videos suggested by someone else via email, blogs, *etc.*; and the remaining 16 respondents (3%) expressed that they could not choose between these options because they did both activities without a clear distinction.

These results reveal a predominant *proactive* behavior by users of video sharing websites. It is confirmed –less strongly however– by Q2 (51% versus 48%).

Additionally, an intuitive observation was also confirmed: *proactive* users access these sites the most frequently (every day). This association was not present in non proactive

[4]Although many subjects started to answer the questionnaires (Q1: 634, Q2: 553), not all of them finished it or complied to the requirements for validating hypothesis 2 and 3 (first: 217, second: 231), as explained in Section 3.3.

Figure 1: Tree graph characterizing the profile of users on video sharing web sites according to the responses obtained with the second questionnaire. Edges width represent the proportion of subjects shared between two nodes.



users, *i.e.,* users that typically watch videos suggested by others ($p < .01$ and $p < .01$ for Q1 and Q2 respectively). Tables 3 and 4 show a cross-tabulation of these variables.

Table 3: Cross-tabulation between variables *find-video* and *video-freq* from Q1 ($C = .20, p < .01$).

Frequency of use of video sharing web sites	Proactive	Passive	Total
Less than once a month	9	13	22
1-3 times a month	9	21	30
1-3 times a week	52	59	111
4-6 times a week	66	51	117
Every day	126	70	196
Total	262	214	476[a]

[a]From 634 subjects, 14 never used video sharing websites, 128 were filtered out[◇], and 16 did not respond about their usual behavior.

Table 4: Cross-tabulation between variables *find-video* and *video-freq* from Q2 ($C = .24, p < .01$)

Frequency of use of video sharing web sites	Proactive	Passive	Total
Less than once a month	7	17	24
1-3 times a month	9	27	36
1-3 times a week	48	68	116
4-6 times a week	79	52	131
Every day	111	71	182
Total	254	235	489[b]

[b]From 553 subjects, 6 never used video sharing websites, 51 were filtered out[◇], and 7 did not respond about their usual behavior.

Table 5: Cross-tabulation between variables *find-video* and *have-account* from Q2 ($\phi = -.006, p = .89$).

Account on a video sharing website	Proactive	Passive	Total
No	92	87	179
Yes	166	153	319
Total	258	240	498[c]

[c]From 553 subjects, 6 never used video sharing websites, 42 were filtered out[◇], and 7 did not respond about their usual behavior.

[◇]The questionnaire was distributed in 9 pages. Among all questions from the first page, we had control questions to characterize how often subjects use video sharing websites, whether they had visual impairments, and whether they could reproduce sounds on their PC. Depending on the answers to those questions we did not display the other pages of the questionnaire.

Q2 also captured which subjects had an account on at least one video sharing website. Interestingly, having an account does not have a significant effect on the user being a proactive or a passive user to video sharing web sites ($\phi = -.006, p = .89$). Table 5 crosses these variables.

q3. *"How many subjects search for specific videos instead of browsing without anything in mind?"* In Q1, from the 266 proactive users of video sharing websites, 168 reported typically searching for specific videos. Additionally, 118 participants out of the 216 passive users stated that although they usually watch videos suggested by others, when they search for videos, they look for something specific. Therefore, 57% of all subjects search for specific videos and are prone to obtain NDVC in the result set of a video search task. With respect to Q2, this fraction was a bit higher (59%). In Q2, we also captured *how* users search for specific videos: (1) typing keywords in the search box, or (2) using the categories available on the main page of a video sharing web site. Results reveal that the majority of subjects (88%) *type keywords* when searching for a specific video. Figure 1 summarizes these findings and characterizes the profiles of the users of video sharing websites. Based on the findings presented herein, we corroborate H1: *Video search is the main method for reaching content on video sharing websites.*

4.2 Validation of H2

> *Identical or approximately identical videos differing in photometric features (image quality), audio quality, editing of the content (i.e., few or more scenes), additional content (i.e., audio and image overlays), or having the same visual context but different audio (or viceversa) are considered by the users as similar clips.*

Next, we present the results obtained about the participants' perception of NDVC when varying the most common low-level features addressed in the literature (see Section 3.2). In addition, the implications of our findings for each variation are discussed with respect to the following variables: (1) differences in image quality, (2) differences in audio quality, (3) differences in visual content, (4) differences in audio content, (5) differences in audio+visual content, and (6) similar semantics on different videos. Tables 6 and 7 summarize the results obtained with Q1 and Q2 respectively.

Differences in image quality (condition A). According to Tables 6 and 7, identical videos with different image quality were perceived as NDVC by both samples in Q1 and Q2 (a majority of 42.9% and 46.8% respectively stated that videos from condition A are "essentially the same"). No significant difference was found between the results from Q1 and Q2 ($p = .10$), thus reinforcing the reliability of the sampling methodology.

Impact of image expertise: In Q2 we asked participants if they considered themselves to be image experts (five-point Likert scale). One could argue that image experts are more sensitive to differences in image quality between two videos. However, this correlation was not significantly different from zero ($\rho = -.03, p = .62$). Table 8 shows a cross-tabulation between the similarity level of the NDVC from condition A and the participants' image expertise.

Differences in audio quality (condition E). Results obtained with Q2 did not clarify whether participants considered NDVC in condition E to be exact duplicates (35.5% of subjects) or near-duplicates (38.1% of subjects). This uncertainty was untied by Q1, as a majority of 45.6% participants considered videos E1 and E2 to be "exactly the same". Although Q1 highlighted this similarity level as the most predominant for condition E, no significant difference was found between results from Q1 and Q2 ($p = .08$). This means that it is not clear whether users perceive NDVC with different audio quality as exactly the same or not. However, this assumption is strengthened by the fact that 41% of the subjects did not notice any change in the audio quality of NDVC from condition E, while only 33% did not notice changes in the image quality related to video clips from condition A. Note that this difference is not due to samples with different levels of image and audio expertise, as no significant difference could be found between these measures ($p = .26$). Given the fact that users perceived NDVC from condition A as essentially the same, these findings support the assumption that *users are more tolerant to changes in the audio than in the video tracks.* Another interesting result was that the level of audio expertise had a significant yet small negative correlation with the similarity attributed to NDVC in condition E ($\rho = -.18, p < .01$). Table 9 shows a cross-tabulation between these measures.

Impact of the audio settings: One could argue that differences in audio quality can be perceived more clearly with headphones than with speakers, which implies that the audio sets of the participants might have affected the decisions (Q1, speakers: $n = 184$, headphones: $n = 33$; Q2, speakers: $n = 159$, headphones: $n = 72$). However, this was not the case ($p = .11$ and $p = .15$ in Q1 and Q2 respectively), meaning that speakers and headphones offered the same similarity level for the musical clips E1 and E2 in both questionnaires.

Tables 10 and 11 show a cross-tabulation between the audio equipment used by participants and the similarity levels attributed to the NDVC from condition E.

Table 6: Similarity levels attributed to each NDVC pair used in Q1 (see Table 2). Figures in bold highlight the highest value for each video pair.

	Video examples (% of subjects)						
Similarity level (five point Likert scale)	A	B	C	D	E	F	G
Completely different	3.2	8.8	5.1	6.0	5.1	2.8	30.0
Essentially different	11.1	14.7	12.9	15.2	9.7	10.6	18.4
Somehow related	7.4	33.2	34.6	23.0	8.3	34.1	**41.9**
Essentially the same	**42.9**	**35.0**	**35.0**	**43.3**	31.3	**45.6**	9.7
Exactly the same	35.5	8.3	12.4	12.4	**45.6**	6.9	0.0

Table 7: Similarity levels attributed to each NDVC pair used in Q2 (see Table 2). Figures in bold highlight the highest value for each video pair.

	Video examples (% of subjects)						
Similarity level (five point Likert scale)	A	B	C	D	E	F	G
Completely different	4.8	13.9	6.9	7.4	3.5	9.5	37.7
Essentially different	13.0	13.9	14.7	18.2	11.7	5.6	15.6
Somehow related	7.4	**39.0**	**40.3**	25.5	11.3	33.8	**39.4**
Essentially the same	**46.8**	27.7	29.9	**39.0**	**38.1**	**47.6**	7.4
Exactly the same	28.1	5.6	8.2	10.0	35.5	3.5	0.0

Table 8: Cross-tabulation between variables *cond-A-similar* and *image-expert* from Q2 ($\rho = -.03, p = .62$)

	Similarity of NDVC in condition A					
*visual expertise	complet. differ-ent	essent. differ-ent	related some-how	essent. the same	exactly the same	total
5	1	4	4	13	14	36
4	4	10	7	37	24	82
3	5	11	5	37	14	72
2	1	5	1	19	9	35
1	0	0	0	2	4	6
total	11	30	17	108	65	231

*1=strongly agree, 2=agree, 3=neither agree, nor disagree, 4=disagree, 5=strongly disagree

Table 9: Cross-tabulation between variables *cond-E-similar* and *audio-expert* from Q2 ($\rho = -.18, p < .01$)

	Similarity of NDVC in condition E					
*audio expertise	complet. differ-ent	essent. differ-ent	related some-how	essent. the same	exactly the same	total
5	2	2	3	2	4	13
4	4	6	1	10	10	31
3	1	13	8	29	25	76
2	0	6	11	30	28	75
1	1	0	3	17	15	36
total	8	27	26	88	82	231

*1=strongly agree, 2=agree, 3=neither agree, nor disagree, 4=disagree, 5=strongly disagree

Table 10: Cross-tabulation between variables *audio-set* and *cond-E-similar* from Q1 ($C = .19$, $p = .11$)

Similarity levels (condition E)	Speakers	Headphones	Total
Completely different	11	0	11
Essentially different	16	5	21
Related somehow	17	1	18
Essentially the same	61	7	68
Exactly the same	79	20	99
Total	184	33	217

Table 11: Cross-tabulation between variables *audio-set* and *cond-E-similar* from Q2 ($C = .17$, $p = .15$)

Similarity levels (condition E)	Speakers	Headphones	Total
Completely different	5	3	8
Essentially different	21	6	27
Related somehow	13	13	26
Essentially the same	59	29	88
Exactly the same	61	21	82
Total	159	72	231

Differences in visual content (condition B). From the results obtained in Q1, no direct conclusion could be drawn on whether participants considered video clips B1 and B2 to be somehow related (33.2% of subjects) or es-

sentially the same (35%). As shown in Table 7, the predominant level of similarity in Q2 was "somehow related" (39% against 27.7% for "essentially the same"). Although the results obtained with both Q1 and Q2 in condition B preserved the same distribution shape and shared most of its properties ($p = .22$, K-S test), there was a significant difference in terms of the median location ($p = .03$, M-W test). In other words, these results basically do not diverge from each other, but Q2 was able to highlight the most probable median. We assume that the presence of additional visual content in one of the videos was the main factor that shifted the users' perception towards a non near-duplicate evaluation.

Differences in audio content (condition D). Condition D uses both audio and visual overlays. However, the analysis of the subjective answers in Q1 revealed that the visual overlay was rarely perceived while the audio overlay characterized the difference between video clips D1 and D2 (D1 was the original video of a plane landing at Bilbao's airport and D2 was the same video with audio comments from a TV newscast and the TV channel's logo at the bottom right side of the screen). That said, the videos were considered to be near-duplicates, as shown in Tables 6 and 7 (majorities of 43.3% and 39% for Q1 and Q2 respectively). In addition, there was no significant difference between the results obtained in each of the questionnaires ($p = .13$), which confirms the reliability of the measure. Given that the videos from condition B were not perceived as near-duplicates, these findings reinforce the assumption that *users are more tolerant to changes in the audio quality than in the video quality.*

Differences in visual+audio contents (condition C). As in condition B, the NDVC from condition C were labeled as "somehow related" (34.6%) or "essentially the same" (35%) in Q1. Once again, the draw was resolved by Q2, where the video clips C1 and C2 were clearly not considered to be near-duplicates (40.3% against 29.9%). Note that the results in Q1 and Q2 preserved the same shape and properties of the distributions ($p = .28$, K-S test). However, Q2 revealed a significant difference in their medians ($p = .04$, M-W test). This means that results from both questionnaires are consistent, but Q2 highlighted the most probable median. Findings from condition C are in agreement with conditions B and D in the sense that additional visual content in each NDVC is an important factor to shift the users' perception towards a non near-duplicate evaluation.

Similar semantics on different videos (conditions F and G). With respect to semantics [1], most subjects perceived videos in condition F as "essentially the same" (45.6% and 47.6% in Q1 and Q2 respectively) and in condition G as "somehow related" (41.9% and 39.4% in Q1 and Q2 respectively). No significant difference was found between the results from Q1 and Q2 for conditions F ($p = .36$) and G ($p = .13$), which enhances the reliability of these results. Note that video clips with different audio and similar visual content (condition F) were considered to be near-duplicates while those with similar audio and different visual content were not (condition G). Again, this observation supports the assumption that users are more tolerant to changes in the audio than in the video channels. Moreover, the semantics between two different videos in condition F led subjects to think of them as NDVC while exact duplicates with overlays in condition D did not. Another interesting result is that only 29% of the subjects considered the changes between NDVC from condition F to be relevant, which was the smallest proportion among all conditions (A: 39%, B: 50%, C: 72%, D: 62%, E: 36%, G: 87%). In other words, two exact duplicates that only differ in their audio or image quality are perceived as having more relevant differences than two totally different videos –with different audio, people, and scenario– that are semantically the same. Therefore, we conclude that the human perception of NDVC has a semantic component. However, it is not clear from our study the exact role that semantics play on particular instances of videos.

Complementary results. In Q2, after evaluating the similarity level of each NDVC pair, participants were asked if: (1) they did not notice any difference between the videos, (2) they noticed differences but did not care about them, or (3) the differences were relevant. There was a strong association between the answers to this question and the answers to the similarity level between NDVC pairs ($p < .01$). This finding reinforces the validity of our experiment and confirms that participants did not respond to the questionnaire randomly.

Conclusion for H2. From the results obtained with our sample, it seems that videos that differ in image quality, audio quality, or audio content (overlay) *are* typically considered to be near-duplicates, in accordance with the literature. However, videos where the transformations applied to the visual content include overlays or insertions of additional information *do not* seem to be perceived as near-duplicates [5]. Furthermore, completely different videos with the same semantics seem to be perceived as near-duplicates, which is not taken into account by most of the definitions in the literature. These observations contradict our hypothesis, and so *we reject H2.* Moreover, an interesting finding is that *users are more tolerant to changes in the audio than in the video tracks.*

4.3 Validation of H3

> *Once the users obtain the result list fo a video search query and after watching the NDVC in such a list, they have a preference for one NDVC over the others and therefore would rather only see the preferred NDVC in the results*

As explained in Section 3, after each similarity evaluation between two NDVC, subjects were asked to report their preferences (if any) about having one/both/none of the videos listed as a result of executing the query search (see Table 2 for information on the queries). Tables 12 and 13 summarize the main results for Q1 and Q2 respectively.

These findings confirm that given two NDVC, users typically prefer to have only one video listed in a video search task, being it the one with: (a) the best image quality (Q1: 52.5%, Q2: 56.7%), (b) the best audio quality (Q1: 35.5%, Q2: 41.1%), or with additional information using (c) overlays (Q1: 46.5%, Q2: 48.9%) or (d) increased length (Q1: 61.3%, Q2: 70.1%). Moreover, participants preferred to have just the original musical clip in condition G instead of both clips.

Conversely, subjects preferred to have both video clips listed when they: (a) shared most scenes but each had additional information (Q1: 53.5%, Q2: 43.3%), or (b) were semantically similar, but visually different (Q1: 44.7%, Q2: 47.2%). In order to understand this behavior, we analyzed all the qualitative answers provided by each participant in Q1. This manual analysis supported our belief that *participants were not able to choose between NDVC that had*

[5]The reader might object that clips of condition D were considered as near-duplicates, even if they presented visual overlay. However, as highlighted previously, we noticed from the analysis of the subjective answers in Q1 that the visual overlay was rarely perceived while the audio overlay characterized the difference between video clips D1 and D2.

Table 12: Preferences over near-duplicates for each NDVC pair used in Q1 (see Table 2). Figures in bold highlight the highest value for each video pair.

Preference (single choice)	Video examples (% of subjects) A	B	C	D	E	F	G
Only video 1	1.8	6.0	5.1	6.0	**35.0**	6.0	**54.4**
Only video 2	**52.5**	14.7	**61.3**	**46.5**	3.2	13.4	6.5
Both videos 1 and 2	18.0	**53.5**	19.4	27.2	24.4	**44.7**	36.4
None of the videos	0.5	4.1	0.5	1.4	1.8	2.3	0.9
No preference	26.3	19.8	13.4	18.4	**35.0**	33.6	1.4
Didn't underst. query	0.9	1.8	0.5	0.5	0.5	0.0	0.5

Table 13: Preferences over near-duplicates for each NDVC pair used in Q2 (see Table 2). Figures in bold highlight the highest value for each video pair.

Preference (single choice)	Video examples (% of subjects) A	B	C	D	E	F	G
Only video 1	1.7	13.4	3.0	8.2	**41.1**	8.2	**59.3**
Only video 2	**56.7**	15.2	**70.1**	**48.9**	5.6	12.1	7.4
Both videos 1 and 2	15.2	**43.3**	18.6	31.6	23.8	**47.2**	28.6
None of the videos	1.7	4.3	0.4	0.4	1.7	2.2	1.3
No preference	24.2	22.5	7.8	10.8	26.8	29.0	2.6
Didn't underst. query	0.4	1.3	0.0	0.0	0.9	1.3	0.9

different pieces of information in them. This assumption holds even for condition F, when participants were focusing on the concept being taught (*i.e.*, atmospheric pressure) instead of the video *per se*. Once again, the results obtained with both Q1 and Q2 did not reveal a significant difference in any of the seven conditions, which ensures the reliability of our findings (A: $p = .68$, B: $p = .10$, C: $p = .23$, D: $p = .14$, E: $p = .38$, F: $p = .46$, G: $p = .55$).

While these preferences are probably video and user dependent, our results certainly give information on how interested people are in having all related video clips listed after executing a query search.

5. IMPLICATIONS FOR DESIGN

The findings of our study have direct implications on the design of retrieval engines for video-sharing websites. Particularly, our results suggest that the way duplicates are treated in the search results should adapt to the feature(s) that make the clips alike.

Note that in our work we have not considered NDVC that infringe copyrights or that maliciously harm the system. With this observation in mind, the core result of our work is that not all near duplicate videos should be treated the same and hence not all should *a priori* be removed from the result list. From the evidence gathered in our study, we propose three features that would improve –from a user-centric perspective– the way search engines treat NDVC: (a) a *user-centric definition of NDVC* that takes into account semantic similarity, (b) a strategy for *clustering the results* around the most representative videos, and a recommendation for (c) *adapting the results* to the specific features that make the clips alike and to the user's video and audio literacy.

5.1 A User-Centric NDVC Definition

Our results suggest that videos that vary in visual content –by overlaying or inserting additional information– were not considered to be near-duplicate of the original videos. Additionally, our results suggest that users of multimedia repositories might benefit from a search engine that takes into account the semantic similarity of the multimedia content. Therefore we propose the following *user-centric definition of NDVC*, which restricts Wu *et al.*'s definition [16] and includes elements of Basharat *at al.*'s definition [1]:

> NDVC are approximately identical videos that might differ in encoding parameters, photometric variations (color, lighting changes), editing operations (captions, or logo insertion), or audio overlays. Conversely, identical videos with relevant complementary information in any of them (changing clip length or scenes) are not considered as NDVC.
>
> Furthermore, users perceive as near-duplicates videos that are not alike but that are visually similar and semantically related. In these videos the same semantic concept must occur without relevant additional information (*i.e.*, the same information is presented under different scene settings).

It must be noted that a fuller user-centric definition of near-duplicate video clips must include the user's intention or interest in the multimedia content. Consider video pair G: one includes the original clip of the Beatles singing "All You Need is Love", the other the same song covered by another band. The audio will be decisive if the user is after the *authentic* version, but not so if the simply want the song in order to learn to play it. Audio will be irrelevant if they are in fact wanting to have a laugh at some 60s (70s?) hairstyles. In our definition *relevance is defined with respect to a goal.* In the presented study, we did not study the interplay of the user's intention and his/her perception of similarity. However, future work should try to refine the proposed definition to incorporate the user's goal(s).

The participants of our study identified clips with the same semantic content as being essentially the same. This result supports research on algorithms to detect semantic similarity, such as the work by Basharat *et al.* [1]. However, the mapping from low-level features onto semantic features is still an open research problem. We believe that this is one of the most promising and challenging research areas in multimedia information retrieval.

5.2 Clustering

The traditional approach to multimedia (images and video) search and retrieval has leveraged the available metadata (tags, comments, surrounding text) in order to compute the similarity between the user-submitted textual query and the content associated to the metadata. Advanced content-based techniques analyze the content of the multimedia material in order to assess the similarity between different items. This is also the case of the NDVC detection algorithms discussed in Section 2 (*e.g.*, [14, 16]).

Given two NDVC, the participants of our study preferred to have only one of the videos listed in the result list of a video search task. Therefore, we propose to use NDVC detection algorithms to create clusters of clips that share video, audio, or semantic content, such that: (1) The clusters would be ranked against the user-submitted query; and (2) only the most representative videos in each cluster would be shown in the result list (cluster centroid). For example, the video to be shown would be the one with the best image or audio quality, or with additional information using overlays, in relation to the results presented in this paper.

A similar attempt was presented by Hsu and colleagues [8]. They proposed an approach for re-ranking search results that preserved the maximal mutual information between the search relevance and the high-dimensional low-level visual features of the videos. However, their approach did not take

into account all the NDVC features tested in the study presented in this paper.

How these clusters are visualized and presented to the user is an open research question. An option would consist of displaying only one representative video per cluster and allowing users to expand the content of the cluster in order to see all duplicate clips belonging to it.

5.3 Feature and User Adaptation of Search Results

Our final recommendation in the design of video retrieval engines consists of adapting the ranking of the results to the features that make clips alike, and to the ability of the user to perceive the differences between the clips.

Our findings support boosting the ranking of NDVC that have more content (*i.e.*, condition C), more information such as subtitles of commentary audio (*i.e.*, condition D), or better video quality (*i.e.*, condition A). In addition, we found significant differences in the perception of NDVC by users with different auditory skills. Therefore and depending on the user's auditory skills, a boost in ranking to clips that have better audio quality might be appropriate (*i.e.*, condition E). Also, video sharing web sites could apply user modeling techniques in order to dynamically update the user's preferences and choose the cluster centroid according to the user's abilities, task and search query.

Further research is required to understand how *simultaneous* differences in more than one feature might interact with the users' perception of similarity. However, we believe that a flexible weighting scheme that would adjust the search results to the specific features of the multimedia content and to the user's abilities would improve user satisfaction with multimedia search engines.

6. CONCLUSIONS AND FUTURE WORK

The findings reported in this paper support the idea that the human perception of NDVC matches many of the features present in its technical definitions with respect to manipulations of non-semantic features [14, 16]. However, similar clips differing in overlaid or added visual content with additional relevant information were not perceived as near-duplicates. Furthermore, we found evidence that users perceive as near-duplicates videos that are not alike but which are visually similar and semantically related (in agreement with Basharat *et al.* [1]).

These findings lead us to propose a user-centric definition of NDVC and a set of user-centric guidelines for the design of video sharing websites. More research is needed to identify low-level features that determine the semantic similarity between two videos. Future work will also include research on the NDVC feature set and psychophysical experiments of the interaction between features.

Acknowledgments

We would like to thank all the participants of our study for their valuable feedback. Also, we would like to thank the anonymous reviewers of this paper for their valuable feedback.

7. REFERENCES

[1] Basharat, A., Zhai, Y., and Shan, M. Content based video matching using spatiotemporal volumes,. *Journal of Computer Vision and Image Understanding 110*, 3 (June 2008), 360–377.

[2] Benevenuto, F., Duarte, F., Rodrigues, T., Almeida, V. A., Almeida, J. M., and Ross, K. W. Understanding video interactions in YouTube. In *Proc. of MM'08* (New York, NY, USA, 2008), ACM, pp. 761–764.

[3] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. of IMC '07* (New York, NY, USA, 2007), ACM, pp. 1–14.

[4] Celebi, M. E., and Aslandogan, Y. A. Human perception-driven, similarity-based access to image databases. In *Proc. of the Artificial Intelligence Research Society Conference* (Clearwater Beach, Florida, May 15–17 2005), I. Russell and Z. Markov, Eds., pp. 245–251.

[5] Gill, P., Li, Z., Arlitt, M., and Mahanti, A. Characterizing users sessions on YouTube. In *Proc. of MMCN'08* (San Jose, CA, USA, January 30-31 2008), ACM.

[6] Guyader, N., Borgne, H. L., Hérault, J., and Guérin-Dugué, A. Towards the introduction of human perception in a natural scene classification system. In *Proc. of Neural Networks for Signal Processing* (Martigny, Switzerland, September 4-6 2002), pp. 385–394.

[7] Halvey, M. J., and Keane, M. T. Exploring social dynamics in online media sharing. In *Proc. of WWW '07* (New York, NY, USA, 2007), ACM, pp. 1273–1274.

[8] Hsu, W. H., Kennedy, L. S., and Chang, S.-F. Video search reranking via information bottleneck principle. In *Proc. of MM'06:* (New York, NY, USA, 2006), ACM, pp. 35–44.

[9] Kruitbosch, G., and Nack, F. Broadcast yourself on YouTube: really? In *Proc. of HCC'08* (New York, NY, USA, 2008), ACM, pp. 7–10.

[10] Maia, M., Almeida, J., and Almeida, V. Identifying user behavior in online social networks. In *Proc. of SocialNets'08* (New York, NY, USA, 2008), ACM, pp. 1–6.

[11] Payne, J. S., and Stonham, T. J. Can texture and image content retrieval methods match human perception? In *Proc. of Intelligent Multimedia, Video and Speech Processing* (Hong Kong, China, 2001), pp. 154–157.

[12] Rui, Y., Huang, T., and Chang, S. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation 10*, 4 (April 1999), 39–62.

[13] Shao, J., Shen, H. T., and Zhou, X. Challenges and techniques for effective and efficient similarity search in large video databases. In *Proc. of the VLDB'08* (2008), vol. 1, pp. 1598–1603.

[14] Shen, H. T., Zhou, X., Huang, Z., Shao, J., and Zhou, X. UQLIPS: a real-time near-duplicate video clip detection system. In *Proc. of VLDB '07* (2007), pp. 1374–1377.

[15] Sheskin, D. J. Handbook of Parametric and Nonparametric Statistical Procedures. Boca Raton, FL: Chapman & Hall/CRC, 2004. 1193 pp.

[16] Wu, X., Hauptmann, A. G., and Ngo, C.-W. Practical elimination of near-duplicates from web video search. In *Proc. of MM'07* (New York, NY, USA, 2007), ACM, pp. 218–227.