

Looking at Near-Duplicate Videos from a Human-Centric Perspective

RODRIGO DE OLIVEIRA, MAURO CHERUBINI, NURIA OLIVER, Telefonica Research, Barcelona

15

Popular content in video sharing websites (e.g., YouTube) is usually replicated via identical copies or near-duplicates. These duplicates are usually studied because they pose a threat to site owners in terms of wasted disk space, or privacy infringements. Furthermore, this content might potentially hinder the users' experience in these websites. The research presented in this article focuses around the central argument that there is no agreement on the technical definition of what these near-duplicates are, and, more importantly, there is no strong evidence that users of video sharing websites would like this content to be removed. Most scholars define near-duplicate video clips (NDVC) by means of non-semantic features (e.g., different image/audio quality), while a few also include semantic features (i.e., different videos of similar content). However, it is unclear what features contribute to the human perception of near-duplicate videos. The findings of four large scale online surveys that were carried out in the context of our research confirm the relevance of both types of features. Some of our findings confirm the adopted definitions of NDVC whereas other findings are surprising: Near-duplicate videos with different image quality, audio quality, or with/without overlays were perceived as NDVC. However, the same could not be verified when videos differed by more than one of these features at the same time. With respect to semantics, it is yet unclear the exact role that it plays in relation to the features that make videos alike. From a user's perspective, participants preferred in most cases 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. Based on all these findings, we propose a new user-centric NDVC definition and present implications for how duplicate content should be dealt with by video sharing Web sites.

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

General Terms: Experimentation, Human Factors, Measurement, Verification

Additional Key Words and Phrases: Psychophysical experiment, similarity, user study, YouTube NDVC, near-duplicate, video sharing

ACM Reference Format:

Oliveira, R., Cherubini, M., and Oliver, N. 2010. Looking at near-duplicate videos from a human-centric perspective. *ACM Trans. Multimedia Comput. Commun. Appl.* 6, 3, Article 15 (August 2010), 22 pages.
DOI = 10.1145/1823746.1823749 <http://doi.acm.org/10.1145/1823746.1823749>

1. INTRODUCTION & MOTIVATION

Today's video sharing websites allow their users to freely post multimedia content without typically checking for its uniqueness. As a consequence, it is not unusual to find in these sites multiple copies

Telefonica Research participates in the Torres Quevedo subprogram (MICINN), cofinanced by the European Social Fund, for researchers recruitment.

Author's address: R. de Oliveira, Via Augusta 177, 09021, Barcelona, Spain; email: oliveira@tid.es.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2010 ACM 1551-6857/2010/08-ART15 \$10.00

DOI 10.1145/1823746.1823749 <http://doi.acm.org/10.1145/1823746.1823749>

ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 6, No. 3, Article 15, Publication date: August 2010.

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 to spam creation [Benevenuto et al. 2008] and copyright infringements [Shen et al. 2007]. For example, Wu et al. [2007] 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 underestimate the role played by NDVC, as they are neither necessarily uploaded with malicious intent nor are exact copies of the original video. In fact, it is not infrequent to find near-duplicate video clips that *complement* the original material with additional information (e.g., commentary audio or subtitles) and that hence might provide valuable information to the users of the system. For instance, a popular YouTube videos might contain news about a recent event, such as the last victory of the Conservative party in the UK elections. However, the audio channel of the original news piece might have been removed to include an audio commentary of the user. This new content might offer the user's perspective on the event and might be considered useful by other users. Hence, it is not clear that all near-duplicate content should automatically be removed from video sharing sites.

In addition, 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 article, we present the results of four 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 Web sites;
- (2) watch pairs of NDVC and state their degree of similarity (some pairs differed in only one feature while others differed in more than one feature); and
- (3) elicit their preferences—if any—on which duplicate they would like to have in the search results

The analysis of the answers to the questionnaires led us to a user-centric definition of NDVC with implications for how duplicated videos should be retrieved in video sharing websites (see Section 6).

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 analyzed data from 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 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

¹See <http://www.youtube.com>, lastly retrieved in March 2010.

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. [2008] 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 antisocial 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 [2007], who analyzed social interactions on YouTube by crawling user pages and focusing on Web-site-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. [2008], 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. [2008] 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*, that is, 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 Web sites.

2.2 Near-Duplicate Video Clips (NDVC)

Turning now our attention to NDVC, it is important to understand the role that duplicated clips play on the way people use video sharing Web sites. In this regard, Kruitbosch and Nack [2008] 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 [Kruitbosch and Nack 2008]. Given that most of the popular content in video sharing Web sites has

Table I. Comparison of NDVC Definitions

Author	NDVC Definition
Wu et al. [2007]	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 insertion), different lengths, and certain modifications (frames add/remove).
Shen et al. [2007]	Clips that are similar to or nearly duplicate of each other, but appear differently due to various changes introduced during capturing time (camera viewpoint and setting, lighting condition, background, foreground, <i>etc.</i>), transformations (video format, frame rate, resize, shift, crop, gamma, contrast, brightness, saturation, blur, age, sharpen, <i>etc.</i>), and editing operations (frame insertion, deletion, swap and content modification).
Basharat et al. [2008]	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.

been found to be professionally generated, one would expect to find a significant number of NDVC in these sites.

Cha et al. [2007] 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. [2007] 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. Therefore multimedia information retrieval scholars have proposed in recent years approaches to detect and cluster this content, in order to eliminate NDVC from the search results.

The first step when building a NDVC detection system is a working definition of NDVC. Table I 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. [2007] tried to identify and remove NDVC using the definition reported in Table I. 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. [2007] extended the definition of NDVC by including changes introduced during capturing time, such as a change of camera viewpoint (see Table I). 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. In 2009, they proposed an enhanced version of their

system that was taking advantage of the user interaction with the system [Cheng et al. 2009]. Yet another definition was employed by Basharat et al. [2008], who included intraclass variations such as scene settings, different viewpoints, different camera motions, to name a few.

More recently, sophisticated techniques for identifying and removing NDVC from large datasets have been proposed. Cheng and Chia [2010] proposed a method of identification and removal of near-duplicates in video search results via matching strata of keyframes. Yang et al. [2009] introduced a system consisting of three major elements: a unified feature to describe both images and videos, a core indexing structure to assure the most frequent data access occurred in the main memory, and a multi-steps verification for queries to best exclude false positives and to increase the precision. Similarly, Zhou et al. [2009] proposed an adaptive frame selection strategy called furthest point Voronoi (FPV) to select the shot frame set according to content and frame distribution. While these previous techniques looked at low-level features of video clips, seek Min et al. [2009] proposed an advanced technique based on identifying semantic concepts along the temporal axis of a particular video sequence, resulting in the construction of a so-called semantic *video signature*. Furthermore, Kim et al. [2010] proposed to use the popular gene sequence alignment algorithm in Biology, that is, BLAST, to detect near-duplicate images.

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:

Ⓕ2, 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.

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:

Ⓕ3, Once the users obtain the result list for 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 search results.

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 [Rui et al. 1999; Shao et al. 2008]: 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). Stating the questions in a neutral manner is also important as people tend to focus on different features when thinking about similarity than when thinking about differences [Tversky 1977]. 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 [Payne and Stonham 2001; Guyader et al. 2002; Celebi and Aslandogan 2005], thus highlighting the importance of extending user studies of human perception to video similarity as well.

3. HYPOTHESES AND APPROACH

In summary, the work presented in this article aims at providing evidence on: (1) how users of video sharing websites 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. The research hypotheses of our work are the following.

- ℍ1. Video search is the main method for reaching content on video sharing websites.
- 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.
- ℍ3. Once the users obtain the result list for 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 search results.

We believe that the validation of these three hypotheses will be instrumental in the development of efficient, useful and intuitive search and retrieval systems of audiovisual content.

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 Web sites (see appendix, section B) 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 (see an example in the appendix, section C), where each pair of NDVC differed in only one feature, as detailed in Section 4.2 (study 1, or *S1*). Subjects were asked to rate the similarity of the paired videos and to state *why* they chose a particular degree of similarity. To study the perception of similarity we removed all possible confounding variables and offered the subjects only a video pair at a time, as suggested by psychological research in the field [Tversky 1977]. In a second study, or *S2*, we edited video pairs that had been considered by the users to be NDVC in such a way that the updated videos would differ in more than one feature, in order to study the interaction between the different low-level features. Moreover, participants were asked to rate the degree of similarity between the edited video pairs and their preferences between videos (study 2, or *S2*).

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" (see Appendix, C).

4. METHODOLOGY

Any video retrieval system whose end user is a human being would greatly benefit from the study of the perception of video content from a psychophysical perspective. In our work, we have conducted a psychophysical experiment² 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 [Payne and Stonham 2001; Guyader et al. 2002; Celebi and Aslandogan 2005]. 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.

²This is the analysis of perceptual processes by studying the effect on a subject's experience or behavior of systematically varying the properties of a stimulus along one or more physical dimensions [Bruce et al. 1996].

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.

In order to test each of the three hypotheses presented in Section 3, we designed 4 large-scale questionnaires ($Q1...Q4$) organized in two studies: $Q1$ and $Q2$ were part of the first study, $S1$, and were deployed in the first quarter of 2009. While $Q1$ presented open-ended questions and served as a testbed for the following questionnaires, $Q2$ presented only multiple-choice questions that were derived from the coding of the open questions of the first questionnaire. $Q3$ and $Q4$ were part of $S2$ and were deployed in the first quarter of 2010. To avoid overwhelming the respondents, the sample of $S2$ was split into $Q3$ and $Q4$, which presented complementary versions of the same questions.

4.1 Procedure

To test our hypotheses, we conducted two studies, each one deploying two large-scale questionnaires on one of the most visited news portals in Spain.³ 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. As an incentive, three 100-euro vouchers were raffled among all respondents. The presence of the incentive did not have any influence on the results collected as participants were informed that they would be included in the lottery mechanism regardless of their specific answers to the questions in the questionnaire. Furthermore, the incentive mechanism we designed in relation to the effort they had to make to complete the questionnaire (i.e., 15 minutes) are standard practices in experiments involving human subjects. 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. This procedure was important to find out which subjects answered the same questionnaire more than once and thus eliminate redundant data. Some overlap between participants of the first and second studies might have happened since we used the same sampling methodology (i.e., recruiting subjects through a news portal). However, the second study was conducted 11 months after the first one, which drastically reduces the influence of residues between them.

In the first study ($S1$), participants watched near-duplicate videos that differed only by one feature (e.g., image quality), while participants from the the second study ($S2$) watched NDVC that differed by more than one feature (e.g., image and audio quality).

The two questionnaires deployed in $S1$ ($Q1$ and $Q2$) had the primary goal of collecting both qualitative and quantitative information from participants while avoiding potential biases in the answers. With respect to $S2$, two questionnaires ($Q3$ and $Q4$) were deployed as a copy of $Q2$, but with different videos presented in each of them. The first questionnaire deployment ($Q1$) lasted one week and the questions related to $\mathbb{H}1$ and to the *why*-component of $\mathbb{H}2$ 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, after participants had defined the similarity between the clips of a particular condition in $Q1$, 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 to the choice: “I noted relevant differences between the videos.”

In order to validate $\mathbb{H}2$, we selected NDVC examples from YouTube following the procedure described in Section 4.2. In $S1$, the presentation order of the seven video examples followed a Latin square design to avoid bias, thus creating seven groups (i.e., ABCDEFG, GABCDEF, FGABCDE, and so forth). Similarly, in $S2$ the presentation order also followed a Latin square design for the three video examples

³See <http://www.terra.es>, lastly retrieved in March 2010.

(i.e., ADE, EAD and DEA). Each participant was submitted randomly to only one group. For each of the seven pairs (conditions) in $S1$ and the three pairs in $S2$, 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.

4.2 Stimuli

In order to validate $\mathbb{H}2$, 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 result set, we identified five NDVC pairs that exemplified variations of the most common non-semantic features [Shen et al. 2007; Wu et al. 2007], and two pairs that illustrated variations of semantic features [Basharat et al. 2008]. The selected videos were edited such that all NDVC pairs would have about the same length ($\bar{x} = 37$ seconds), except in condition C (see Table II).⁴ These video pairs deployed in questionnaires $Q1$ and $Q2$ differed by only one feature at a time (e.g., image quality).

After analyzing data from $Q1$ and $Q2$, we edited video pairs that users considered to be NDVC in such a way that the updated videos differed by two or three features at the same time. Therefore, we created $Q3$ and $Q4$ as a replica of $Q2$ but with different sets of video pairs. In the former, we presented videos maximizing differences (e.g., video $A1_{Q3}$ with worse image and audio quality and video $A2_{Q3}$ with better image and audio quality), while in the latter we included videos balancing differences (e.g., video $A1_{Q4}$ with worse image quality and better audio quality, and video $A2_{Q4}$ with better image quality and worse audio quality). Table II provides some information about each video pair presented in $S2$.

4.3 Participants

An initial pool of 2496 participants answered part of the questionnaires from both studies $S1$ and $S2$. Both samples had subjects with a wide range of occupations. In terms of validating $\mathbb{H}1$, we considered only subjects that complied with the following requirements: (1) fluent in Spanish; (2) had experience with at least one video-sharing Web site; (3) answered all questions about their use of video-sharing Web sites; (4) could listen to the audio track in the videos by means of the computer speakers or a headphone; and (5) had no significant audio or video impairment. Therefore, a total of 1335 respondents were considered in this data analysis ($Q1$: 313 subjects; $Q2$: 304 subjects; $Q3$: 356 subjects; $Q4$: 362 subjects).

In terms of validating $\mathbb{H}2$ and $\mathbb{H}3$, we postfiltered the $\mathbb{H}1$ sample to consider only subjects that: (1) spent at least the minimum amount of time to fill out the questionnaires and watch their videos.⁵ and (2) provided answers to all of the questions related to the videos. Furthermore, in the analysis of each of the four questionnaires, we considered the same amount of subjects per group in terms of the presentation order of the video examples (see Section 4.1). That said, a total of 448 respondents of

⁴The clips used in $S1$ can be viewed at: <http://goo.gl/BYhb>, while the clips used in $S2$ can be viewed at: <http://goo.gl/9UxC>. Last retrieved in June 2010.

⁵Subjects took medians of 18 and 19 minutes to answer questionnaires $Q1$ and $Q2$ respectively, and a median of 9 minutes to answer questionnaires $Q3$ and $Q4$. As 8.7 minutes are required to watch the 14 videos (7 NDVC pairs) in questionnaires $Q1$ and $Q2$, we stipulated 10 minutes as the minimum to answer each of them. Similarly, we considered 5 minutes as the minimum duration time to answer questionnaires $Q3$ and $Q4$, given that at least 3.3 minutes are necessary to watch the 6 videos (3 NDVC pairs) presented in each of them.

Table II. Descriptions of the Videos Used in the Four Questionnaires



























Q	Condition	Query	Video 1	Video 2
Q1, Q2	A Photometric variation	crazy frog champions	A1: standard image 	A2: higher quality (better colorfulness and lighting) 
	B Edt. op. (add/rmv 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 Edt. op. (AV 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 img. and diff. 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 
Q3	A Photometric variation	crazy frog champions	A3: better image and audio quality 	A4: worse image and audio quality 
	D Edt. op. (AV overlays)	plane airport Bilbao wind	D3: better image and audio quality 	D4: worse image and audio quality (audio comments and logo) 
	E Audio quality	More than Words	E3: better image and audio quality (44KHz) 	E4: worse image and audio quality (mono audio in 11KHz) 
Q4	A Photometric variation	crazy frog champions	A5: better image and worse audio quality 	A6: worse image and better audio quality 
	D Edt. op. (AV overlays)	plane airport Bilbao wind	D5: better image and worse audio quality 	D6: worse image and better audio quality (comments and logo) 
	E Audio quality	More than Words	E5: better image and worse audio q. (44KHz) 	E4: worse image and better audio q. (mono 11KHz) 

Table III. Descriptive Statistics of the Participants of Studies $S1$ and $S2$

	Study $S1$ (February 2009)		Study $S2$ (March 2010)	
	$Q1$	$Q2$	$Q3$	$Q4$
Initial pool of subjects	634	553	668	641
Valid answers for $\mathbb{H}1^\circ$	313 (m: 164)	304 (m: 173)	356 (m: 216)	362 (m: 248)
Age (mean)	31.4 ($s = 8.83$)	33.4 ($s = 9$)	32.8 ($s = 10.38$)	32.9 ($s = 9.69$)
Computer usage (median)*	5 (iqr = 0)	5 (iqr = 0)	5 (iqr = 0)	5 (iqr = 0)
Video sharing usage (median)*	4 (iqr = 2)	4 (iqr = 2)	4 (iqr = 1)	5 (iqr = 1)
Audio expertise (median)**	—	3 (iqr=1)	3 (iqr = 1)	3 (iqr = 1)
Image expertise (median)**	—	2 (iqr=1)	3 (iqr = 1)	3 (iqr = 1)
Valid answers for $\mathbb{H}2$ and $\mathbb{H}3^\bullet$	217 (m: 105)	231 (m: 136)	159 (m: 105)	165 (m: 117)
Age (mean)	31.5 ($s = 9.05$)	33.2 ($s = 9.11$)	32.4 ($s = 9.5$)	31 ($s = 9.28$)
Computer usage (median)*	5 (iqr = 0)	5 (iqr = 0)	5 (iqr = 0)	5 (iqr = 0)
Video sharing usage (median)*	4 (iqr = 2)	4 (iqr = 2)	4 (iqr = 2)	5 (iqr = 1)
Audio expertise (median)**	—	3 (iqr = 1)	3 (iqr = 2)	3 (iqr = 1)
Image expertise (median)**	—	2 (iqr = 1)	3 (iqr = 2)	3 (iqr = 1)

$^\circ$ Subjects that (1) were fluent in Spanish; (2) had experience with at least one video-sharing Web site; (3) answered all questions related to how they use video-sharing Web sites; (4) could listen to the audio track in the videos by means of the computer speakers or a headphone; (5) had no significant audio or video impairment.

$^\bullet$ Subjects that (1) followed the restrictions imposed for the $\mathbb{H}1$ validation; (2) spent at least the minimum amount of time possible to fill out the questionnaire; and (3) answered all questions related to the videos. Each of the four questionnaires preserved the same amount of subjects per group regarding the presentation order of the video examples.

*5-point scale: 1: less than once a month; 2: 1–3 times a month; 3: 1–3 times a week; 4: 4–6 times a week; 5: everyday.

**5-point scale: 1: totally disagree that I am an expert; 5: totally agree that I am an expert.

$Q1$ and $Q2$ were taken into account to validate $\mathbb{H}2$ and $\mathbb{H}3$ when near-duplicates differed by only one feature. In order to understand the implications of this validation when the NDVC differed by more than one feature, we analyzed the answers to $Q3$ and $Q4$ (324 subjects). Table III summarizes the profile of the participants recruited for both studies.

4.4 Measures

Multiple choice questions with a single answer were used to test both $\mathbb{H}1$ and $\mathbb{H}3$, whereas $\mathbb{H}2$ 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.

4.5 Statistical Analysis

Dependent variables from each questionnaire were either nominal (e.g., strategy to search videos, preferred NDVC) or ordinal (e.g., frequency of computer usage, similarity between NDVC). Therefore, we opted for a non-parametric approach to: 1) highlight differences between variables and 2) calculate associations and correlations between them. With respect to the first goal, the Kolmogorov-Smirnov test (K-S test) and the Mann-Whitney U test (M-W test) were used to identify differences between two independent samples at the ordinal level (e.g., similarity of videos A1 and A2 in $Q1$ and $Q2$). Similarly but for nominal variables, we used the Chi-square test (χ^2) to verify differences between distributions (e.g., preference between videos A1 and A2 in $Q1$ and $Q2$). With respect to the second goal, other statistics derived from the Pearson Chi-Square were used, such as the Phi coefficient (ϕ) to measure the association between two dichotomies (e.g., v1: *find-video*—whether participants watch videos found by themselves or suggested by someone else; v2: *have-account*—whether users have or don't 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*; $v2$: *video-freq*—how frequently subjects use video sharing websites). Finally, the Spearman’s Rho (ρ) was used to measure correlations between two ordinal related variables (e.g., $v1$: similarity level between videos A1 and A2; $v2$: subject’s image expertise).

5. RESULTS AND DISCUSSION

5.1 Validation of $\mathbb{H}1$

Video search is the main method for reaching content on video sharing Web sites.

The following methodology was used to falsify this hypothesis. First, we identified how many participants of the four questionnaires use video-sharing Web sites *proactively* (q1): that is, when they watch a video, it is usually a video that they found by themselves instead of being suggested by someone else. Second, we highlighted the fraction of these participants that usually have a *purpose* when searching for specific videos instead of browsing with nothing in mind (q2). 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 Web sites, we reject the hypothesis. An example of a proactive search would be a user looking for videos of a specific song from “The Beatles,” going to a video sharing Web site, typing the title of the song in the search box, and going through the results.

q1. “How many subjects use video sharing Web sites proactively?” From the 1335 participants of the four questionnaires that answered all questions necessary to validate $\mathbb{H}1$, 786 (59%) reported watching videos found by themselves; 522 (or 39%) reported watching videos suggested by someone else via email, blogs, etc.; and the remaining 27 respondents (2%) 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 Web sites.

Interestingly, having an account on at least one video-sharing Web site has a weak association with the user’s proactive attitude to search videos on these Web sites ($N = 1335$, $\chi = 11.195$, $\phi = -.092$, $p < .01$). Furthermore, having an account does not have a significant effect on how frequently users watch videos on these Web sites ($N = 1335$, $\rho = -.048$, $p = .08$).

q2. “How many subjects search for specific videos instead of browsing without anything in mind?” From the 786 proactive users of video-sharing Web sites, 492 reported typically searching for specific videos. Additionally, 289 participants out of the 522 passive users stated that although they usually watch videos suggested by others, when they search for videos, they look for something specific. Therefore, 59% 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 questionnaires $Q2$, $Q3$ and $Q4$, 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 website. Results reveal that the majority of subjects (91%) *type keywords* when searching for a specific video. Based on the findings presented in this section, we corroborate $\mathbb{H}1$.

5.2 Validation of $\mathbb{H}2$ (Part 1: Videos Differing by Only One Feature)

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.

Table IV. Similarity Levels Atributed to Each NDVC Pair Used in Q1 and Q2 (see Table II). Figures in Bold Highlight the Highest Value for Each Video Pair

Similarity level (5-point scale)	Conditions in questionnaire Q1							Conditions in questionnaire Q2						
	A	B	C	D	E	F	G	A	B	C	D	E	F	G
Completely different	3.2	8.8	5.1	6.0	5.1	2.8	30.0	4.8	13.9	6.9	7.4	3.5	9.5	37.7
Essentially different	11.1	14.7	12.9	15.2	9.7	10.6	18.4	13.0	13.9	14.7	18.2	11.7	5.6	15.6
Somehow related	7.4	33.2	34.6	23.0	8.3	34.1	41.9	7.4	39.0	40.3	25.5	11.3	33.8	39.4
Essentially the same	42.9	35.0	35.0	43.3	31.3	45.6	9.7	46.8	27.7	29.9	39.0	38.1	47.6	7.4
Exactly the same	35.5	8.3	12.4	12.4	45.6	6.9	0.0	28.1	5.6	8.2	10.0	35.5	3.5	0.0

Table V. Cross-Tabulation between Variables *Cond-A-Similar* and *Image-Expert* from Q2 ($\rho = -.03$, $p = .62$)

*visual Expertise	Similarity of NDVC in Condition A					Total
	Compleat. Different	Essent. Different	Related Somehow	Essent. The Same	Exactly The Same	
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.

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 4.2). 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. Table IV summarizes the results obtained with questionnaires Q1 and Q2.

Differences in image quality (condition A). According to Table IV, 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 V 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 nearly the same. 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

Table VI. Cross-Tabulation between Variables *Cond-E-Similar* and *Audio-Expert* from Q2 ($\rho = -.18, p < .01$)

*Audio Expertise	Similarity of NDVC in Condition A					Total
	Compleat. Different	Essent. Different	Related Somehow	Essent. The Same	Exactly The Same	
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 VII. Cross-Tabulation between Variables *Audio-Set* and *Cond-E-Similar* from Q1 ($C = .19, p = .11$) and Q2 ($C = .17, p = .15$)

Similarity levels (cond. E)	Q1			Q2		
	Speakers	Headphones	Total	Speakers	Headphones	Total
Completely different	11	0	11	5	3	8
Essentially different	16	5	21	21	6	27
Related somehow	17	1	18	13	13	26
Essentially the same	61	7	68	59	29	88
Exactly the same	79	20	99	61	21	82
Total	184	33	217	159	72	231

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 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 VI 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.

Table VII shows a cross-tabulation between the audio equipment used by participants and the similarity levels attributed to the NDVC from condition E.

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 essentially the same (35%). As shown in Table IV, 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 perception.

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 Table IV (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 [Basharat et al. 2008], 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 image or audio quality (conditions A and E respectively) are perceived as having more relevant differences than two different videos—with different audio, people, and scenario—that are semantically the same (condition F). 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. Significant correlations between the answers to this question and to the similarity level of NDVC pairs could be observed under different levels: weak (B: $\rho = -.27$), moderate (C: $\rho = -.32$, D: $\rho = -.33$, F: $\rho = -.41$, G: $\rho = -.30$) and strong (A: $\rho = -.53$, E: $\rho = -.72$). This finding reveals a somewhat obvious finding: the more users perceive video pairs as similar videos, the less significant differences they find between them. This observation reinforces the validity of our experiment and confirms that participants did not respond to the questionnaire randomly.

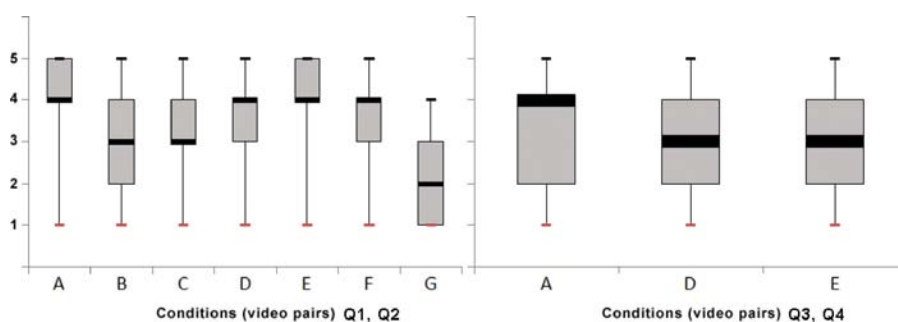


Fig. 1. (left) Boxplot of similarity of video pairs from questionnaires $Q1$ and $Q2$. (right) Boxplot of similarity of video pairs from questionnaires $Q3$ and $Q4$. Note that the interquartile ranges in all conditions reveal that videos could be considered as essentially *different*.

Preliminary conclusion for $\mathbb{H}2$. From the results obtained with questionnaires $Q1$ and $Q2$, duplicate videos that differ in image quality (condition A), audio quality (condition E) or with/without overlays (condition D) are considered to be near-duplicates (similarity level in conditions A and E: $\tilde{x} = 4, q1 = 4, q3 = 5$; and D: $\tilde{x} = 4, q1 = 3, q3 = 4$). Conversely, videos with different audio or visual content were not considered NDVC (similarity level in conditions B: $\tilde{x} = 3, q1 = 2, q3 = 4$; and C: $\tilde{x} = 3, q1 = 3, q3 = 4$). Furthermore, completely different videos with the same semantics seem to be perceived as near-duplicates (similarity level in condition F: $\tilde{x} = 4, q1 = 3, q3 = 4$), which is not taken into account by most of the definitions in the literature. Figure 1 presents this information visually by aggregating the results from both questionnaires $Q1$ and $Q2$.

Although these preliminary results already contradict hypothesis $\mathbb{H}2$, part of the NDVC technical definition remains accurate (i.e., for videos differing in image quality, audio quality, or with/without overlays). Therefore, the following subsection discusses the results of questionnaires $Q3$ and $Q4$, in which videos differ by more than one feature.

5.3 Validation of $\mathbb{H}2$ (Part 2: Videos Differing by More than One Feature)

As described in Section 4.2, questionnaires $Q3$ and $Q4$ contained the same questions as those from $Q2$, but including only three of the original video pairs, that is, those from conditions A, D and E. These videos were chosen because in study $S1$ users considered them to be near-duplicates. Therefore, in $S2$ we wanted to investigate whether the same perception would be maintained when these videos differed by more than one feature at a time, as suggested by the NDVC technical definition. In order to compare data between studies $S1$ and $S2$, first we looked for differences between $Q1$ and $Q2$, and between $Q3$ and $Q4$, regarding similarity between videos from conditions A, D and E. Given that no significant difference was found in neither case (in $S1$, A: $p = .10$; D: $p = .13$; E: $p = .08$; and in $S2$, A: $p = .63$; D: $p = .62$; E: $p = .45$), we could make a straight comparison of similarity between the video pairs in $S1$ and $S2$. This comparison revealed that each of the three conditions provided a different perception in $S2$ when compared to $S1$ ($p < .01$). The main outcome of this analysis is that near-duplicate videos differing by more than one feature are considered to be less similar than near-duplicate videos differing by only one feature. Moreover, videos differing by more than one feature were actually not considered to be near-duplicates, as depicted by Figure 1.

After taking a closer look at the results from $S2$, we observed that the users' perception of similarity between videos followed a bimodal distribution (see Figure 2). This characterizes a divergence of opinion among participants, who could either consider near-duplicate videos differing by more than

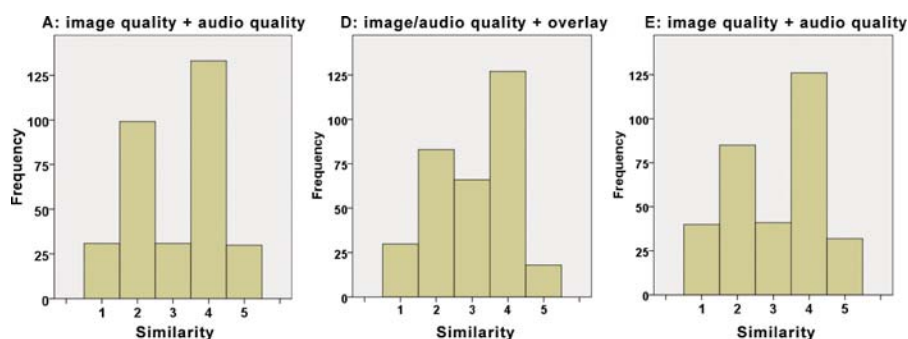


Fig. 2. Similarity of video pairs in conditions A, D and E (study $S2$). The bi-modal distributions reveal that users were uncertain whether videos differing by more than one feature at a time could be considered as near-duplicates or videos that are essentially different.

one feature as “essentially the same” or “essentially *different*.” Interestingly, this finding refutes the current technical definition of NDVC.

Another interesting finding is related to the subjects’ perception of relevant differences in video pairs. In questionnaires $Q3$ and $Q4$, we asked users that found relevant differences between videos to tell us what were the main differences. Answers were later manually categorized into: (1) differences in image quality, (2) differences in audio quality, and (3) differences in audio/image content, that is, insertion of audio/visual overlays. Whenever users mentioned more than one of these categories, each one was counted once for the same user. Results from both questionnaires indicate that users consider either audio quality or image quality as the most relevant difference between videos that differ by these two features at the same time. While in condition A users were more concerned with image quality than audio quality ($Q3$: 103 vs. 52 votes for video and audio respectively; $Q4$: 96 vs. 53 votes respectively), this behavior was inverted in condition E ($Q3$: 52 vs. 83 votes for video and audio respectively; $Q4$: 35 vs. 99 votes respectively). We interpret this phenomenon as follows: Although conditions A and E presented video clips of songs, the video clip in condition A was very colorful while the video clip in condition E was in black and white. Therefore, we assume that image quality was the main feature taken into account by participants in their analysis of how similar were the videos that differed by image and audio quality. This assumption is confirmed by the results from $Q4$. While in condition A there was a clear preference for the video with best image quality and worse audio quality (60%)—instead of the video with better audio and worse image quality (7%), the opposite preference was not as evident in condition E (38% of the respondents opted for the video with better audio quality while 23% preferred the video with better black and white image). These findings further support observations from study $S1$, thus allowing us to conclude that users are more tolerant to changes in the audio than in the video tracks.

Summary and final conclusion for $\mathbb{H}2$. Results from $S1$ revealed that duplicate videos differing by either audio or video content (i.e., editing, different length) are not considered by users to be near-duplicates, which contradicts the NDVC technical definition. Conversely, duplicate videos with different image quality, audio quality or with/without overlays are perceived as NDVC. However, when these videos differed by more than one of these features (study $S2$), users did not perceive them as near-duplicates anymore. Actually, a bimodal distribution was observed in the similarity between video pairs, thus leading to the conclusion that users can either perceive these videos as NDVC or videos that are essentially different. Therefore and considering results from both $S1$ and $S2$, we reject $\mathbb{H}2$.

Table VIII. Preferences Over Near-Duplicates for Each NDVC Pair Used in Q1 and Q2 (See Table II). Figures in BoldHighlight the Highest Value for Each Video Pair

Preference (single choice)	Conditions in Questionnaire Q1							Conditions in Questionnaire Q2						
	A	B	C	D	E	F	G	A	B	C	D	E	F	G
Only video 1	1.8	6.0	5.1	6.0	35.0	6.0	54.4	1.7	13.4	3.0	8.2	41.1	8.2	59.3
Only video 2	52.5	14.7	61.3	46.5	3.2	13.4	6.5	56.7	15.2	70.1	48.9	5.6	12.1	7.4
Both videos	18.0	53.5	19.4	27.2	24.4	44.7	36.4	15.2	43.3	18.6	31.6	23.8	47.2	28.6
Neither videos	0.5	4.1	0.5	1.4	1.8	2.3	0.9	1.7	4.3	0.4	0.4	1.7	2.2	1.3
No preference	26.3	19.8	13.4	18.4	35.0	33.6	1.4	24.2	22.5	7.8	10.8	26.8	29.0	2.6
Didn't underst. query	0.9	1.8	0.5	0.5	0.5	0.0	0.5	0.4	1.3	0.0	0.0	0.9	1.3	0.9

Table IX. Preferences Over Near-Duplicates for Each NDVC Pair Used in Q3 and Q4 (See Table II). Figures in BoldHighlight the Highest Value for Each Video Pair

Preference (single choice)	Conditions in Q3			Conditions in Q4		
	A	D	E	A	D	E
Only video 1	3.8	3.8	69.2	7.3	26.1	23.0
Only video 2	68.6	50.3	3.1	60.0	35.2	37.6
Both videos 1 and 2	15.7	34.0	16.4	20.0	30.3	23.6
Neither videos	0.6	1.3	1.3	1.2	0.0	2.4
No preference	9.4	10.1	8.8	10.3	8.5	12.7
Didn't understand query	1.9	0.6	1.3	1.2	0.0	0.6

5.4 Validation of H3

Once the users obtain the result list for 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 4, 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 II for information on the queries). Tables VIII and IX summarize the main results of the four questionnaires.

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 the best image quality (Q1: 52.5%, Q2: 56.7%), the best audio quality (Q1: 35%, Q2: 41.1%), with additional information by means of overlays (Q1: 46.5%, Q2: 48.9%) or 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 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: $p = .55$).

With respect to video pairs that differed by more than one feature, participants also preferred to have only one video listed in the video search list, namely the one with better image and audio quality

(A_{Q3} : 68.6%, A_{Q4} : 60%, E_{Q3} : 69.2%, E_{Q4} : 37.6%), or with additional information (D_{Q3} : 50.3%, D_{Q4} : 35.2%).

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. That said and considering the results from studies $S1$ and $S2$, we corroborate $H3$.

6. 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 search 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.

6.1 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 the one given by Wu et al. [2007] and includes elements of Basharat et al. [2008]:

Human perception of similarity between video clips is increased by proximity of low-level features, and by semantic relatedness. At the same time, perception of similarity is also diminished by interaction of simultaneous changes in multiple features, and by increased informative value. Furthermore, the perception of similarity is a function of these elements and the context in which the videos are appraised (i.e., the user’s background and intentions).

In other words, 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. However, combinations of these variations can reduce similarity between NDVC to the point of being considered to be different videos. The same occurs for identical videos with relevant complementary information in any of them (changing clip length or scenes). Furthermore, users perceive as near-duplicates videos that are not alike but are visually similar and semantically related. In these videos, the same semantic concept must be present 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 more than attributes inherent to a video—or even its semantics, such as the context of the user and their intention—or lack of one. The bimodal picture found in Figure 2 argues that in some (many?) cases, the definition of near-duplicates cannot be pinned on the videos themselves, but lies with the user, including his/her own experiences and personality, and the intent with which (s)he is browsing. In

other words, it is a function of *video1+video2+user+situation*, not just *video1+video2*. Consider video pair G: one includes the original clip of the Beatles singing “All You Need is Love” and the other contains the same song covered by another band. The audio will be decisive if the user is after the *authentic* version, but not so much if s/he simply wants the song in order to learn how to play it or remember the lyrics. Audio will be irrelevant if users are in fact wanting to have a laugh at some 70s hairstyles. In our definition, relevance is defined with respect to a goal. In the presented study, we did not look at 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 Basharat et al. [2008]. 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.

6.2 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. Sophisticated 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 [Shen et al. 2007; Wu et al. 2007; Yang et al. 2009; Zhou et al. 2009; Cheng and Chia 2010].

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 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 article.

A similar attempt was presented by Hsu et al. [2006]. 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 article.

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.

6.3 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

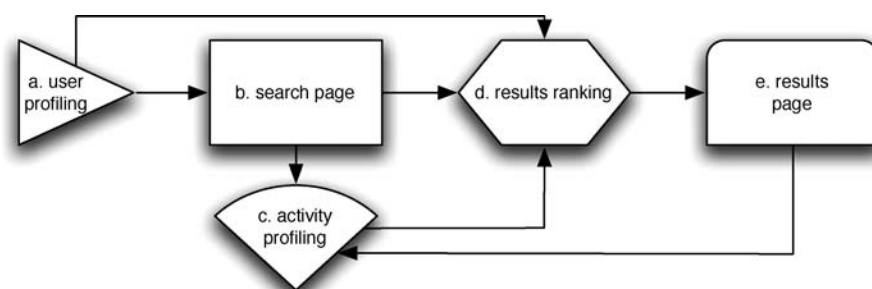


Fig. 3. Interaction flow of an improved social video portal search engine.

websites 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.

6.4 Summary: an Improved Social Video Portal Search Engine

To summarize the implications of this work, we present the elements of a hypothetical search engine of an improved video sharing website. The results of this work suggest that traditional search engines could be enriched by two elements: a user profiling module (part a of Figure 3), and an activity profiling module (part c of Figure 3). While the former creates and maintains an accurate model of the user that might be transversal to different search sessions (e.g., age, gender, group affiliation, auditory skills, etc.), the latter creates and maintains a model of the user's intention in a particular search session (e.g., the user is looking for the original video, else the user search re-edited clips with additional content, the user is looking for—semantically—related videos, the user is getting frustrated, etc.). The activity profiling is built using information that might be explicitly provided by the user in the search page (part b of Figure 3) and implicitly inferred logging the user's behavior while s/he browses the search results (part e of Figure 3). Finally, the outputs of the user profiling and the activity profiling modules could feed the results ranking algorithm (part d of Figure 3).

7. CONCLUSIONS AND FUTURE WORK

The findings reported in this article support the idea that the human perception of NDVC matches many of the features that are already considered in the technical definitions with respect to manipulations of nonsemantic features [Shen et al. 2007; Wu et al. 2007]. However, near-duplicate videos differing by more than one feature at the same time or with/without extra relevant information were not perceived to be near-duplicates in our study. Furthermore, we found evidence that users perceive as near-duplicates those videos that are not alike but that are visually similar and semantically related (in agreement with Basharat et al. [2008]).

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 on our side will include

research on the relation between the user's intention and his/her perception of similarity of NDVC. We are currently designing an improved search mechanism for video sharing websites like the one described in Section 6.4 for a major social network portal in Spain.⁶ The challenges related to this are related to defining and testing multiple rating schemes that could combine the multiplicity of factors described in this article and that could—at the same time—optimally satisfy the user's needs.

ACKNOWLEDGMENTS

We would like to thank the reviewers of this article and all the participants in our study for their valuable feedback. Also, we would like to thank the staff of Terra.es for providing support to the experiments.

REFERENCES

- BASHARAT, A., ZHAI, Y., AND SHAN, M. 2008. Content based video matching using spatiotemporal volumes. *J. Comput. Vis. Image Under.* 110, 3, 360–377.
- BENEVENUTO, F., DUARTE, F., RODRIGUES, T., ALMEIDA, V. A., ALMEIDA, J. M., AND ROSS, K. W. 2008. Understanding video interactions in youtube. In *Proceeding of the 16th ACM International Conference on Multimedia (MM'08)*. ACM, New York, 761–764.
- BRUCE, B., GREEN, P. R., AND GEORGESON, M. A. 1996. *Visual Perception*. 3rd Ed. Psychology Press.
- CELEBI, M. E. AND ASLANDOGAN, Y. A. 2005. Human perception-driven, similarity-based access to image databases. In *Proceedings of the 18th International Florida Artificial Intelligence Research Society Conference*. I. Russell and Z. Markov, Eds. 245–251.
- CHA, M., KWAK, H., RODRIGUEZ, P., AHN, Y.-Y., AND MOON, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC'07)*. ACM, New York, 1–14.
- CHENG, R., HUANG, Z., SHEN, H. T., AND ZHOU, X. 2009. Interactive near-duplicate video retrieval and detection. In *Proceedings of the 17th ACM International Conference on Multimedia (MM'09)*. ACM, New York, 1001–1002.
- CHENG, X. AND CHIA, L.-T. 2010. Stratification-based keyframe cliques for removal of near-duplicates in video search results. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR'10)*. ACM, New York, 313–322.
- GILL, P., LI, Z., ARLITT, M., AND MAHANTI, A. 2008. Characterizing users sessions on youtube. In *Proceedings of the SPIE/ACM Conference on Multimedia Computing and Networking (MMCN)*.
- GUYADER, N., BORGNE, H. L., HÉRAULT, J., AND GUÉRIN-DUGUÉ, A. 2002. Towards the introduction of human perception in a natural scene classification system. In *Proceedings of Neural Networks for Signal Processing*. 385–394.
- HALVEY, M. J. AND KEANE, M. T. 2007. Exploring social dynamics in online media sharing. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, New York, 1273–1274.
- HSU, W. H., KENNEDY, L. S., AND CHANG, S.-F. 2006. Video search reranking via information bottleneck principle. In *Proceedings of the 14th Annual ACM International Conference on Multimedia (MULTIMEDIA'06)*. ACM, New York, 35–44.
- KRUITBOSCH, G. AND NACK, F. 2008. Broadcast yourself on youtube: really? In *Proceeding of the 3rd ACM International Workshop on Human-Centered Computing (HCC'08)*. ACM, New York, 7–10.
- MAIA, M., ALMEIDA, J., AND ALMEIDA, V. 2008. Identifying user behavior in online social networks. In *Proceedings of the 1st Workshop on Social Network Systems (SocialNets'08)*. ACM, New York, 1–6.
- PAYNE, J. S. AND STONHAM, T. J. 2001. Can texture and image content retrieval methods match human perception? In *Proceedings of Intelligent Multimedia, Video and Speech Processing*. 154–157.
- RUI, Y., HUANG, T., AND CHANG, S. 1999. Image retrieval: current techniques, promising directions and open issues. *J. Vis. Comm. Image Repres.* 10, 4, 39–62.
- SEOK MIN, H., CHOI, J., NEVE, W. D., AND RO, Y. M. 2009. Near-duplicate video detection using temporal patterns of semantic concepts. In *Proceedings of the International Symposium on Multimedia*, 65–71.
- SHAO, J., SHEN, H. T., AND ZHOU, X. 2008. Challenges and techniques for effective and efficient similarity search in large video databases. *Proc. VLDB Endow.* 1, 2, 1598–1603.

⁶See <http://www.keteke.es>, last retrieved May 2010.

- SHEN, H. T., ZHOU, X., HUANG, Z., SHAO, J., AND ZHOU, X. 2007. Uqlips: a real-time near-duplicate video clip detection system. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)*. VLDB Endowment, 1374–1377.
- KIM, H.-S., CHANG, H.-W., LEE, J., AND LEE, D. 2010. Effective near-duplicate image detection using gene sequence alignment. In *Advances in Information Retrieval, Lecture Notes in Computer Science*, vol. 5993. Springer, Berlin, 229–240.
- TVERSKY, A. 1977. Features of similarity. *Psych. Rev.* 84, 4, 327–352.
- WU, X., HAUPTMANN, A. G., AND NGO, C.-W. 2007. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA'07)*. ACM, New York, 218–227.
- YANG, X., ZHU, Q., AND CHENG, K.-T. 2009. Near-duplicate detection for images and videos. In *Proceedings of the 1st ACM Workshop on Large-Scale Multimedia Retrieval and Mining (LS-MMRM'09)*. ACM, New York, 73–80.
- ZHOU, X., ZHOU, X., CHEN, L., BOUGUETTAYA, A., XIAO, N., AND TAYLOR, J. A. 2009. An efficient near-duplicate video shot detection method using shot-based interest points. *Trans. Multimedia.* 11, 5, 879–891.

Received March 2010; revised May 2010; accepted June 2010