Human Perception of Near-Duplicate Videos

Rodrigo de Oliveira, Mauro Cherubini and Nuria Oliver

Telefonica Research, Via Augusta, 177
08021 Barcelona, Spain
{oliveira, mauro, nuriao}@tid.es

Abstract. Popular content in video sharing websites (*e.g.,* YouTube) contains many duplicates. Most scholars define near-duplicate video clips (NDVC) as identical videos with variations on non-semantic features (*e.g.,* image/audio quality), while a few others also include semantic features (different videos of similar content). However, it is unclear what exact features contribute to human perception of similar videos. In this paper, we present the results of a user study conducted with 217 users of video sharing websites. Findings confirm the relevance of both classes of features, but the exact role played by semantics on each instance of NDVC is still an open question. In most cases, participants had a preference for one video when compared to its NDVC and they were more tolerant to changes in the audio than in the video channel.

Keywords: NDVC, near-duplicate, similarity, user study, YouTube.

1 Introduction

In the last few years, different research groups have tried to understand how video-sharing web sites are used and in particular the impact that near-duplicate video clips (NDVC) have on video information retrieval tasks [4], spam creation [1] and identification of copyright infringements [2]. Most of the previous work has focused on identifying and removing NDVC. However, there is no agreement on the technical definition of what features identify almost identical copies of the same video. According to Wu and colleagues [4], near duplicate videos differ in file format or encoding parameters, photometric variations (*e.g.*, color) and editing operations (*e.g.*, overlay, captions, add/remove scenes). Other scholars have employed an extended definition of *similarity between features* including changes on capturing time [2], for instance a different camera viewpoint. Finally, Basharat *et al.* included *similarity at the semantic level* [1], where the same semantic concept can occur under different scene settings (*e.g.*, two videos of different deer in different forests grazing moss).

We believe these studies are important but could benefit from additional information gathered via user studies for at least three reasons: 1) little is known about how users are affected by the presence of NDVC; 2) we have limited knowledge of whether the definition of near duplicate videos we choose to adopt matches the users' understanding of what NDVC are; and 3) we need empirical proofs that removing NDVC from the results set of a video search task would satisfy the users' needs. Therefore, we state the hypotheses of our study as:

H1 Users of video sharing websites search for videos more than browse.
H2 Users of video sharing websites perceive NDVC according to their definition [1, 3, 4].
H3 Users of video sharing websites have preferences over near-duplicate videos and usually don't want to have all of them listed and displayed after executing a search query.

To test these hypotheses, we deployed a large-scale qualitative questionnaire where respondents characterized their common use of video sharing websites, watched pairs of NDVC and stated their similarity degree (pairs differing by only one feature), and presented their preferences (if any) about which duplicate they would like to have in the search results. These measurements led us to a user-centered definition of NDVC.

2 Experiment Design

Procedure. A questionnaire was deployed on a popular news portal in Spain (www.terra.es) to test our hypotheses. 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 if they usually use services like YouTube to: (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 typically: (1) found by themselves, (2) suggested by someone else, or (3) found by other means. All subjective answers were manually categorized.

In order to validate H2, we looked for the most viewed videos in YouTube from "last month" and "at all times", and created queries to retrieve these videos. From the results set, five NDVC pairs were identified to exemplify variations of non-semantic features [3, 4], and two pairs to exemplify variations of semantic features [1]. Videos were edited such that all pairs would have the same length ($\bar{x} = 37$ seconds), except in one condition (see Table 1). Video examples were presented on a Latin square basis to avoid bias, thus creating seven groups. Each participant was submitted randomly to only one group. For each of the seven pairs, participants were asked to fully watch both videos and rate how similar they thought these videos were.

Table 1. Descriptions of the NDVC pairs used in the study (http://tinyurl.com/youtubestudy).

Condition	Query	Video 1	Video 2
A: Photometric variation	crazy frog champions	**A1:** standard image	**A2:** higher quality (color 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, 44Khz	**E2:** mono, 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

H3 was addressed by asking participants if they had a preference between the videos belonging to the same condition and which one would it be if they were searching for videos using the same query (see Table 1 for conditions and queries).

Participants. From an initial pool of 647 participants who self-volunteered to answer the questionnaire, 217 (115 male, 122 female) complied with all the study requirements: (1) fluent in Spanish, (2) experience with at least one video sharing website, (3) could listen to the videos using the computer speakers or headphones, and (4) no relevant audio or visual impairment. Their median age was 31 years (min: 16, max: 63), 98.2% were Spanish, and had a wide range of occupations. Subjects reported using computers everyday and video sharing websites from 4 to 6 days/week.

3 Results and Discussion

Validation of H1: *Search was confirmed to be the users' purpose when using video sharing websites.* Sixty-one percent (133 subjects) of participants declared using these systems to search for a *specific* video while 39% (84 subjects) use them to spend time without anything in mind. Moreover, participants were *active users of these systems:* the majority of participants declared watching videos found by themselves (53%).

Validation of H2: Table 2 summarizes the results obtained in terms of H2, where the figure in bold reflects the highest value for each video pair.

Table 2. Similarity levels attributed to each NDVC pair (see Table 1).

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

According to Table 2, identical videos with different image quality (condition A) were perceived as near-duplicates (42.9% stated that videos A1 and A2 are "essentially the same"). Interestingly, when identical videos differed only in audio quality (condition E), they were considered as "exactly the same" (45.6%). One could argue that differences in audio quality could be perceived easier with headphones than with speakers, which suggests that the participants' different audio sets affected decisions (speakers: 184; headphones: 33). However, this was not the case (p=0.11). Therefore, we conclude that in the context of video sharing websites, *users are more tolerant to changes in the audio than in the video modality.* Regarding the validation of H2, given that NDVC from conditions B (add/remove scenes), C (different lengths) and D (overlays) were also mostly rated as "essentially the same", we could corroborate that the human perception of NDVC matches the commonly adopted definitions associated to non-semantic features [3, 4]. However, it is important to note that participants were undecided whether videos from conditions B and C could also be considered as "related somehow" (33.2% *vs.* 35% and 34.6% *vs.* 35% respectively). We are investigating this tie effect on a further quantitative study. With respect to semantics [1], most subjects perceived videos in condition F as "essentially the same" and in condition G as "related somehow". Therefore, we conclude that the *human perception of NDVC also has a semantic component*. However, it is not clear from our study the exact role that semantics play on particular instances of videos.

Validation of H3. Our findings confirm that *given 2 NDVC, users usually prefer one of them in a video search task*, being it the one with best image quality (condition A) or additional information (conditions C and D). In the case of videos sharing audio semantics, they opted for the original musical clip (condition G). Subjects preferred both videos when they shared most scenes with additional information on each (condition B), or were semantically similar, but visually different (condition F). When videos differed in audio quality, subjects either had no preference or preferred the one with the best quality (condition E). Table 3 summarizes these results.

Table 3. Preferences over near-duplicates for each NDVC pair (see Table 1).

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 understand query	0.9	1.8	0.5	0.5	0.5	0.0	0.5

Implications. Near-duplicate detection algorithms could achieve better results by also comparing semantics between video clips (*e.g.*, similar scenes, same people or objects, *etc.*). Furthermore, near-duplicates in search results should be treated according to what feature(s) make clips alike, *e.g.*, when a NDVC has additional information (condition C), its relevance in the search results should be increased.

4 Conclusions

From our results, human perception of NDVC matches many of the features present in its technical definitions with respect to manipulations of non-semantic features [2,4]. However, it is yet not clear whether similar clips differing in overlaid or added visual content with additional information can be considered as near-duplicates. Furthermore, the definition should be extended to videos with similar semantics but different visual and audio information [1]. However, there is still the need to identify low-level features that influence semantic similarity. Results of a follow-up questionnaire are being analyzed to clarify these findings. We plan to extend research on the feature set and include psychophysical experiments of feature interaction.

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. 360–377, 2008.
2. Benevenuto, F., Duarte, F., Rodrigues, T., Almeida, V. A., Almeida, J. M., and Ross, K. W. Understanding video interactions in YouTube. In *MM '08*. New York, USA. ACM, pp. 761–764, 2008.
3. Shen, H. T., Zhou, X., Huang, Z., Shao, J., Zhou, X. UQLIPS: a real-time near-duplicate video clip detection system. In *VLDB'07*. VLDB Endowment, pp. 1374–1377, 2007.
4. Wu, X., Hauptmann, A. G., and Ngo, C.-W. Practical elimination of near-duplicates from web video search. In *MULTIMEDIA'07*. New York, USA. ACM, pp. 218–227, 2007.