# A walk through the web's video clips

Sara Zanetti[†] Lihi Zelnik-Manor[‡] Pietro Perona[†]
[†] California Institute of Technology [‡] Technion
Pasadena, CA 91125, USA Haifa, 32000, Israel
sara.zanetti@gmail.com lihi@ee.technion.ac.il perona@caltech.edu

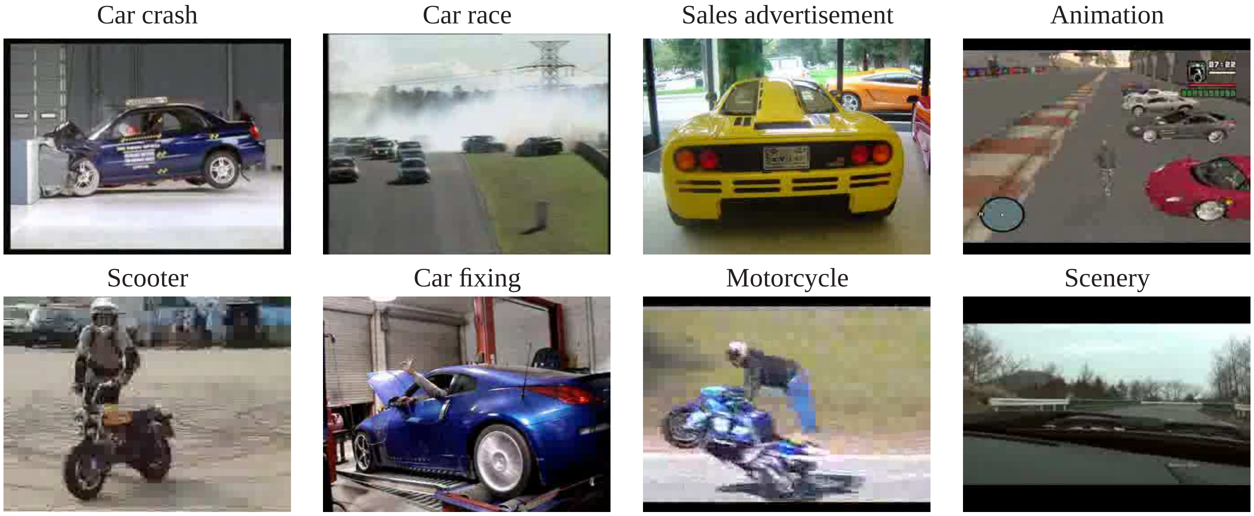

Figure 1. **Data diversity.** Middle frames of 10 video clips from the Autos & Vehicles category show the diversity one finds within category. The category includes clips of car crashes, various races, car fixing instructions, computer animation, and more.

## Abstract

*Approximately $10^5$ video clips are posted every day on the web. The popularity of web-based video databases poses a number of challenges to machine vision scientists: how do we organize, index and search such large wealth of data? Content-based video search and classification have been proposed in the literature and applied successfully to analyzing movies, TV broadcasts and lab-made videos. We explore the performance of some of these algorithms on a large data-set of approximately 3000 videos. We collected our data-set directly from the web minimizing bias for content or quality, way so as to have a faithful representation of the statistics of this medium. We find that the algorithms that we have come to trust do not work well on video clips, because their quality is lower and their subject is more varied. We will make the data publicly available to encourage further research.*

## 1. Introduction

Content based video analysis has attracted extensive research ranging from shot detection [12, 9, 7] to video retrieval [17] and classification [15, 18]. Entire conference series have been dedicated to the topic of content-based retrieval [11]. While in previous years the available video data was either movies, TV broadcasts or home brewed videos, recently the world has seen a revolution in video usage. Web sites such as YouTube [2] made video sharing an immensely popular medium. The goal of this paper is similar to that of TRECVID [11]: we wish to explore and encourage the development of content based analysis of video. We differ, in that we focus on 'typical' web video data. We show that organizing web video databases is still a challenge for machine vision.

Very little work has been done on analysis of the web's video clips. Loui et al. [6] collected a database of a few thousands consumer video clips. Most of the videos were obtained directly from consumers and part were downloaded from YouTube. The collection process included a manual verification step were difficult videos that do not match a set of predefined categories were discarded. The choice of categories was careful according to concepts which are potentially distinguishable. Moreover, the collection was limited to consumer videos and thus more challenging videos such as cartoons or mixture of professional and consumer videos were avoided.

Ulges et al. [14] collected a database of 1650 videos according to 22 predefined categories. The selected categories were again highly different from each other, for example, under the travel and places category the selected tags were: beach, desert, eiffeltower, hiking and sailing, which are typically characterized by different appearances as they are filmed at different sceneries. Appearance based classification into categories of these two databases produced promising results [14, 4]. The authors imply that content based web video search is already plausible.

In this paper we challenge these results. Rather than collecting a database of highly different categories, our goal is to collect a database which represents well the main categories one finds on the web and the search tasks people apply to it. Web video databases can be browsed in two fashions. One can use the division into categories provided by the hosting web site. Alternatively, the database can be searched for keyed-in tags. We therefore collected data from YouTube [2] having both types of browsing in mind. Our database includes the top videos (at the day of collection) of the 11 YouTube [2] categories. It also includes three categories of typical searches: Actors & Actresses, Animals, and Places & Landmarks, where about 10 tags were selected under each of these categories. This data collection captures exactly those difficulties rejected by [6] and [14].

We have collected a database of 2795 video clips and show that classification tests similar to those presented in [14, 4] perform poorly on our data. We further explore the characteristics of web videos and investigate what causes these failures. While [14, 4] used only appearance related features to represent video content, we incorporate also temporal information and show this improves the classification results. Finally, we point at what needs to be done to resolve the situation and allow handling real search tasks.

We start by presenting our data collection methodology in Section 2. We then explore the characteristic of web video data in Section 3 and show that previous algorithms perform poorly in basic tasks Section 4. We conclude in Section 5.

## 2. Collecting a video clip database

We designed a process for data collection that is both practical and minimizes experimenters's bias. Our goal was collecting a set of movies that would be maximally representative of the statistics that one finds on the web.

### 2.1. Technical details about collecting the data

We selected YouTube as our source of video clips since it is currently the most popular site. Video clips were downloaded using the semi-automatic software Ares Tube [1]. We used MPEG4v2 compression with spatial resolution of $320 \times 240$. Over a time period of 6 months we collected a total of 2795 video clips amounting to 15GB of storage space. Downloading time varied depending on web traffic and video files size, but was typically between 20sec and 40sec for 1 minute of video. For most of the data the original URL's were saved, therefore, the data can be downloaded independently by others, avoiding copyright issues.

### 2.2. The methodology of data collection

Browsing through YouTube video clips can be done in two ways: searching for keywords, or sorting by popularity, categories, channels or communities. Both keyword search and category based browsing provide good candidates for data collection and thus we followed both tracks.

- **YouTube Categories.** YouTube sorts its video clip collection into 11 categories. We downloaded $\sim$100 video clips from each category, taking the top video clips of each category without any pruning.
- **Tag search.** We selected three types of popular search topics and further chose about 10 tags under each category. Our goal was to download 100 clips for each tag. However, this task is not yet complete. We watched the first 3-4 seconds of each clip before downloading in order to verify correspondence with the tag – this step does not introduce a selection bias in terms of the quality of the video, nor in terms of which tags are used.

The final collection of video clips is summarized in the following list:

- **11 YouTube categories**:
  - Autos & Vehicles (99 videos), Comedy (92), Entertainment (93), Film & Animation (61), Gadgets & Games (97), Howto & DIY (86), Music (90), News & Politics (93), Pets & Animals (90), Sports (86), Travel & Places (91).
- **Tag search**:
  - **Actors & Actresses** (10 subcategories): Angelina jolie (32), Brad Pitt (49), George Clooney (51), Jennifer Aniston (39), Jim Carrey (101), Julia Roberts (53), Leonardo DiCaprio (93), Penelope Cruz (39), Sharon Stone (34), Tom Cruise (100).

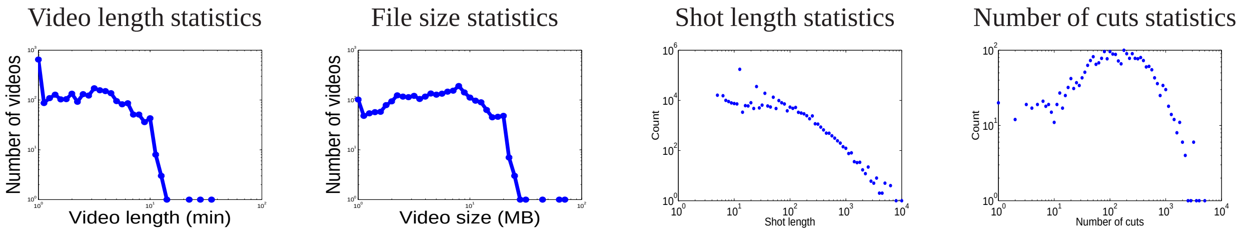

Figure 2. **Web video statistics.** Most web video clips are short (less than 5 min). The number of cuts and the shot length are not useful in distinguishing between categories since the variance within category is large.

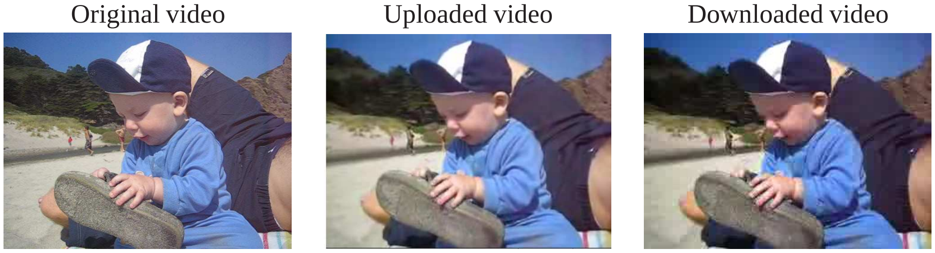

Figure 3. **Degradation in quality.** Uploading and downloading videos degrades their quality significantly and modifies colors. The main degradation in quality occurs at the upload process as can be seen by comparing the original frame on the left with the corresponding screen capture of the uploaded video in the middle. The download process introduces further degradation.

- **Animals** (10 subcategories): Birds (100), Cats (100), Dogs (100), Elephants (49), Fishes (100), Horse (100), Monkeys (23), Sharks (53), Snakes (58), Tigers (15).
- **Places & Landmarks** (8 subcategories): Hong Kong & Central Plaza & Bank of China (56), London & Big Bang (98), New York & Empire State Building (98), Paris & Eiffel Tower (98), San Francisco (60), Roma & Colosseum (99), Toronto (5), Sydney (13).

The download process was not always successful. A large number of clips (between 25%-30%) resulted in corrupt files or illegal structure, such as varying frame dimension within the video. These clips were eliminated.

# 3. Web videos characteristics

Prior to analyzing the data, we explore some basic characteristics of the video clips we collected.

## 3.1. Diversity of data

As YouTube is a video sharing website where (almost) anyone can upload, view and share video clips, the variability among clips is huge. They include the familiar TV shows and movies, professional advertisements and music clips, news broadcasts and documentaries. They also include videos of presentation slides, animation, cartoons, lectures and amateur video which can be any crazy stuff. Figure 1 displays the diversity one finds within a single category.

## 3.2. Basic statistics

The statistics of the collected video clips are summarized in Fig. 2. The average video clip length is 173 sec, the longest clip is 34 min long and the shortest is 5 sec. The mean number of cuts is 227 with a standard deviation of 300, i.e., not informative at all. Further dissecting the data according to categories reveals that there is barely any correlation between video category and video length or shot statistics (see Fig. 2).

## 3.3. Degradation in quality

An additional difficulty when processing web videos is their poor quality. Uploaded videos are down-sampled and their quality is reduced by the host web site. Moreover, the download process often introduces further degradation in quality. An example is presented in Fig. 3. The degradation in quality by upload process is unavoidable. The degradation in the download process depends on the particular implementation and thus could possibly be reduced.

# 4. Do known algorithms work for web videos?

Similar to [6] and [14] we wish to explore the performance of known algorithm on this data. The visual information we extract is based on four components: the cuts one finds in the video, the colors, spatial appearance, and temporal changes (motion). This is since, for example, videos taken by amateurs could imply fewer cuts compared to professional video, and more camera shake. There could be moments of fast motion (when the recording person spins

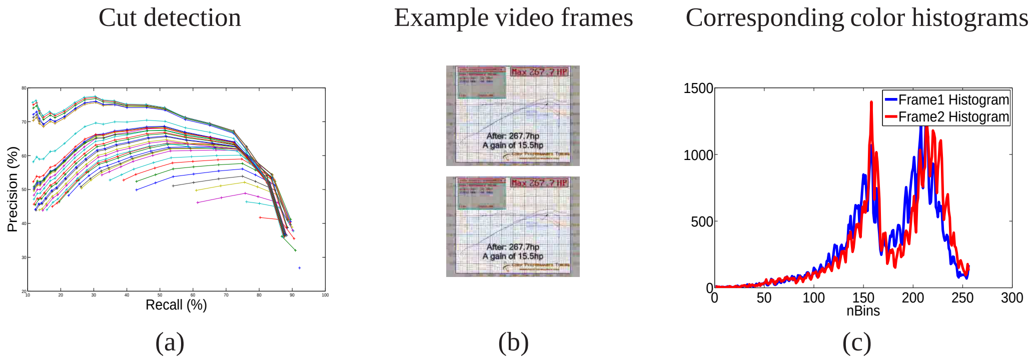

Figure 4. **Poor cut detection.** (a) Cut detection precision-recall values of the algorithm in [12] for a range of threshold values. None are particularly good. (b) The color histograms of consecutive frames are often partially shifted due to poor video quality, even though the frames are highly similar. See text for further details.

around to change view point without shutting the camera off). Sports videos (e.g., tennis, soccer, swimming) could have prevalent saturated background colors and slow panning camera motions. Music videos could have frequent cuts and a certain palette of colors.

While both [6] and [14] experimented with classification only, we start by evaluating the performance of cut detection algorithms. This is since one could expect to find some correlation between cut statistics and video categories. Moreover, cut detection is considered a much simpler and better studied task compared to classification.

### 4.1. Cut detection

We call 'cut' the point where a 'shot' in a movie ends and another begins. Sometimes cuts are abrupt and sometimes one shot fades into another. The mid-point of the brief 'fade sequence' may be taken as a fiducial location of the cut. Detecting cuts and parsing movies into constituent shots is often an important step in video and movie analysis. We have studied the comparison between shot cut detection in [7] and have implemented the algorithm in [12]. We selected at random 5 video clips of 10 different categories to a total of 50 videos and manually labeled all the cuts. We then tested the performance of the shot detection algorithm of [12] which is based on $\chi^2$ distances between consecutive frame color histograms on temporal windows of 10 frames. Their approach uses two threshold, one is used to detect fades and the other to detect cuts. Experimenting with a wide range of these parameters produced results of significantly lower quality compared to the previously published ones; see Fig. 4.a. To boost performance we tried replacing the $\chi^2$ distance measure with more sophisticated distance measures between histograms [10, 5], however, results were not improved. This is due to a number of reasons (see Fig. 4):

- Often the color distribution of frames changes abruptly within the video, even though the scene has not changed (see Fig. 4.b). Since shot detection algorithms rely mostly on color changes these are wrongly detected as cuts. The change in color distribution could be a product of either the original capturing process by a low-quality camera or due to the degradation in quality when uploading and downloading files from YouTube.

- With some surprise we discovered that in approximately 5% of the video clips we downloaded the size of the frames is not constant, and thus the frames include black borders (typically two horizontal stripes) whose width may change. Furthermore, the intensity value of these borders may also change (e.g., from 0 to 1) through the video and thus confuse the cut detection algorithm.

- Cartoons and presentation slides do not comply to the underlying assumptions of shot-cut detection and indeed result in poor performance.

- Severe camera shake and fast motions are often confused as shot boundaries.

Some of these issues could be resolved, e.g., identifying explicitly that a given video includes black borders, these could be eliminated from the video when searching for cuts. Each issue will require a specialized treatment implying a more complex system.

### 4.2. Supervised classification

Searching and indexing of databases require classification capabilities. We have experimented with two supervised classification approaches: Fisher Linear Discriminant (FLD), and K-Nearest Neighbors (KNN). Both gave comparable results and thus we report only those of the latter. The KNN classification was performed in a leave-one-out manner. That is, each video was classified using the rest of the database. For each video we find 11 nearest neighbors

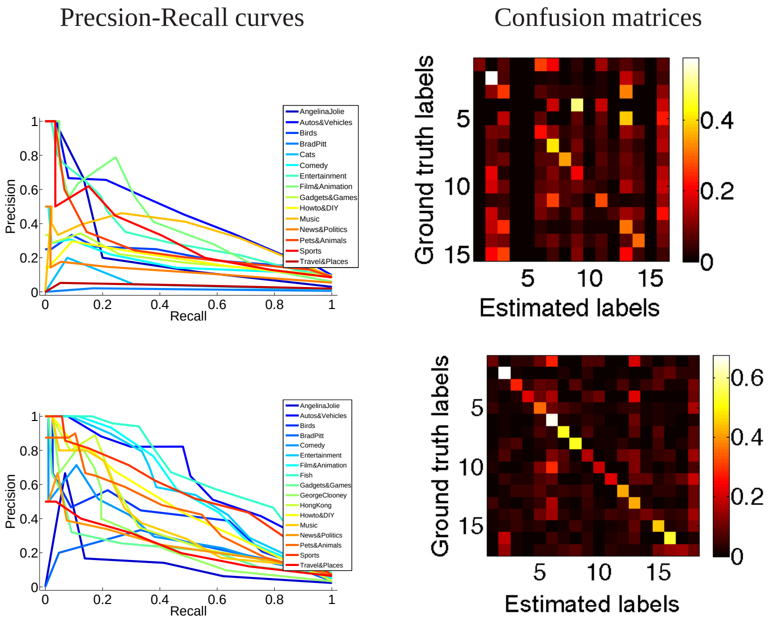

Figure 5. **Supervised classification.** Top: using histograms of visual words [13]. Bottom: using spatio-temporal gradients histograms [16]. Incorporating temporal information is highly important and improves results significantly.

and classify according to the majority vote. If no category receives more than 3 votes, then we assigned the video to an ‘unknown’ category.

We tested five different features as representatives of the video content:

1. Number of cuts. Cuts were detected as described in Section 4.1.

2. Color histogram of the entire video.

3. Histogram of frame-to-frame color changes. We compute the $\chi^2$ distances between color histograms of consecutive frames. We then construct a 3-bin histogram of the $\chi^2$ values, capturing the frequencies of abrupt, medium and slow changes in the video.

4. Histogram of visual words [13].

5. Histogram of muti-scale space-time normalized gradients [16].

The performance of the first three was very low and therefore results are omitted from this paper. The fourth feature is equivalent to the visual features used in [6, 14] and includes spatial appearance information only, ignoring the temporal information in the video. Therefore, we have also tested with space-time gradients [16] and indeed obtained better results, as is shown next.

We performed two sets of experiments. First, repeating the procedure reported in [6, 14], a subset of video clips was selected such that videos from different categories looked different to the experimenter. Classification using either space-time gradient histograms or visual words histograms successfully classified the videos. Therefore, we confirm that in carefully chosen collections of videos, relatively simple classification techniques work well.

We then proceeded to applying the same scheme to the entire unbiased data-set, i.e. 2795 videos divided in 29 categories . The performance degraded significantly to an average correct classification rate of 34% when using KNN classification on the space-time gradient histograms. Results on a subset of 17 categories are shown in Fig. 5. We show only partial results since the processing of the entire database did not finish on time for submission. We observe that (a) classification results are mediocre at best, at least if one uses the tags and class labels as ground truth (we will see later that this may not be a wise choice for testing algorithms). (b) Exploring the precision-recall curves reveals that at low recall (e.g. 0.1) a majority of categories has precision rates above 0.8. This is good news for web-based searches where the raw material is plentiful, and therefore high precision at low recall rates is already useful. Further exploring various combinations of the above listed features provided minimal improvement in reesults.

We believe there are two main reasons for this failure in classification. The first is that the data is too “complex”, as was shown in Section 3.1. In the following section we further strengthen this claim by showing poor unsupervised classification results. The second major difficulty arises from the ground-truth information we have. The assignment into categories in YouTube, as well as the attachment of tags is done by the owner who uploads the clip. These are thus highly subjective and not at all consistent. For example, a video of a car race could be placed under both Autos & Vehicles category and Sports category. The decision is almost random. Moreover, often more than one label matches the same clip, e.g., a clip could be labeled 'soccer', 'funny' and 'cartoon' at the same time. The standard KNN and FLD classification methods we used rely on a single label per data item and are thus inappropriate for the task at hand. To illustrate that we present a few example results of successful and unsuccessful nearest neighbor queries in Fig. 6. In many cases videos from different categories look highly similar, while videos from a single category could seem different.

### 4.3. Unsupervised clustering

Since manual data labeling is not straightforward we further experimented with unsupervised clustering. We tried two popular methods: k-means clustering [3] , which is robust but is limited to blob like clusters, and spectral clustering [8], which can handle complex cluster shapes. Both methods were tested on both the visual words representation and the saptio-temporal gradients histogram representation and provided generally poor results, see Fig. 7. This implies that the underlying categorical structures are highly complex and cannot be discovered by these approaches.

## 5. Summary: What's next?

The experiments of Section 4 show that algorithms developed with nice-and-clean videos in mind, generally do not perform well on web video clips. So, is there any hope for future success? We believe there is. Our study brings to the spotlights two main issues that need to be addressed to allow progress:

- A major difficulty we encountered is in hand labeling the data. Unlike images that can be labeled quickly and easily labeling videos requires numerous man-hours. Tools for video summarization and for fast browsing need to be developed.
- Most videos fit more than one label. We thus need to develop clustering and classification algorithms which allow assigning a video to more than one group, both when hand labeling as well as in the automatic classification.
- Current supervised classification algorithms assume the provided labels are correct and complete. The classification of web video clips require allowing incomplete labeling and considering errors in labeling.

In writing this paper we hope to bring these issues to the research community. We further intend to make the dataset we collected, as well as our baseline results using basic algorithms, publicly available for research purposes.

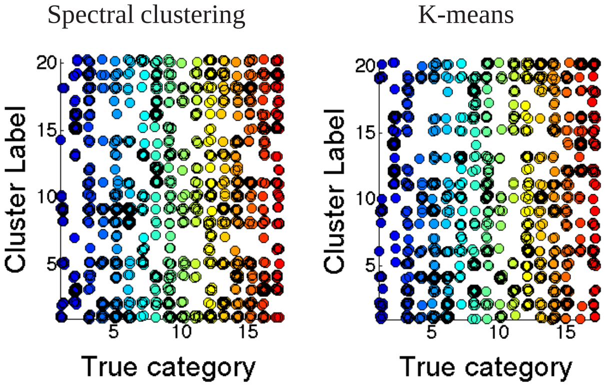

Figure 7. **Unsupervised classification.** using spatio-temporal gradients histograms [16] as features. Both spectral clustering as well as k-means clustering fail to capture the underlying categorical structure.

## References

[1] Ares Tube *http://www.officialares.com/downloads/arestube* 2

[2] YouTube *http://www.youtube.com/* 1, 2

[3] C.M. Bishop, "Neural Networks for Pattern Recognition", Oxford, England: Oxford University Press, 1995. 6

[4] S. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A.C. Loui, and J. Luo. 2007. "Large-scale multimodal semantic concept detection for consumer video". *In Proceedings of the international Workshop on Workshop on Multimedia information Retrieval* (Augsburg, Bavaria, Germany, September 24 - 29, 2007). MIR '07. ACM, New York, NY, 255-264. 2

[5] K. Grauman and T. Darrell. "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features" Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA 4

[6] A., Loui, J., Luo, S., Chang, D., Ellis, W., Jiang, L., Kennedy, K., Lee, and A. Yanagawa, 2007. "Kodak's consumer video benchmark data set: concept definition and annotation". *In Proceedings of the international Workshop on Workshop on Multimedia information Retrieval* (Augsburg, Bavaria, Germany, September 24 - 29, 2007). MIR '07. ACM, New York, NY, 245-254. DOI= http://doi.acm.org/10.1145/1290082.1290117 1, 2, 3, 4, 5

[7] G. Lupatini and C. Saraceno and R. Leonardi,1998. "Scene Break Detection: A Comparison", *In Proceedings of the Workshop on Research Issues in Database Engineering* (February 23 - 24, 1998). RIDE. IEEE Computer Society, Washington, DC, 34. 1, 4

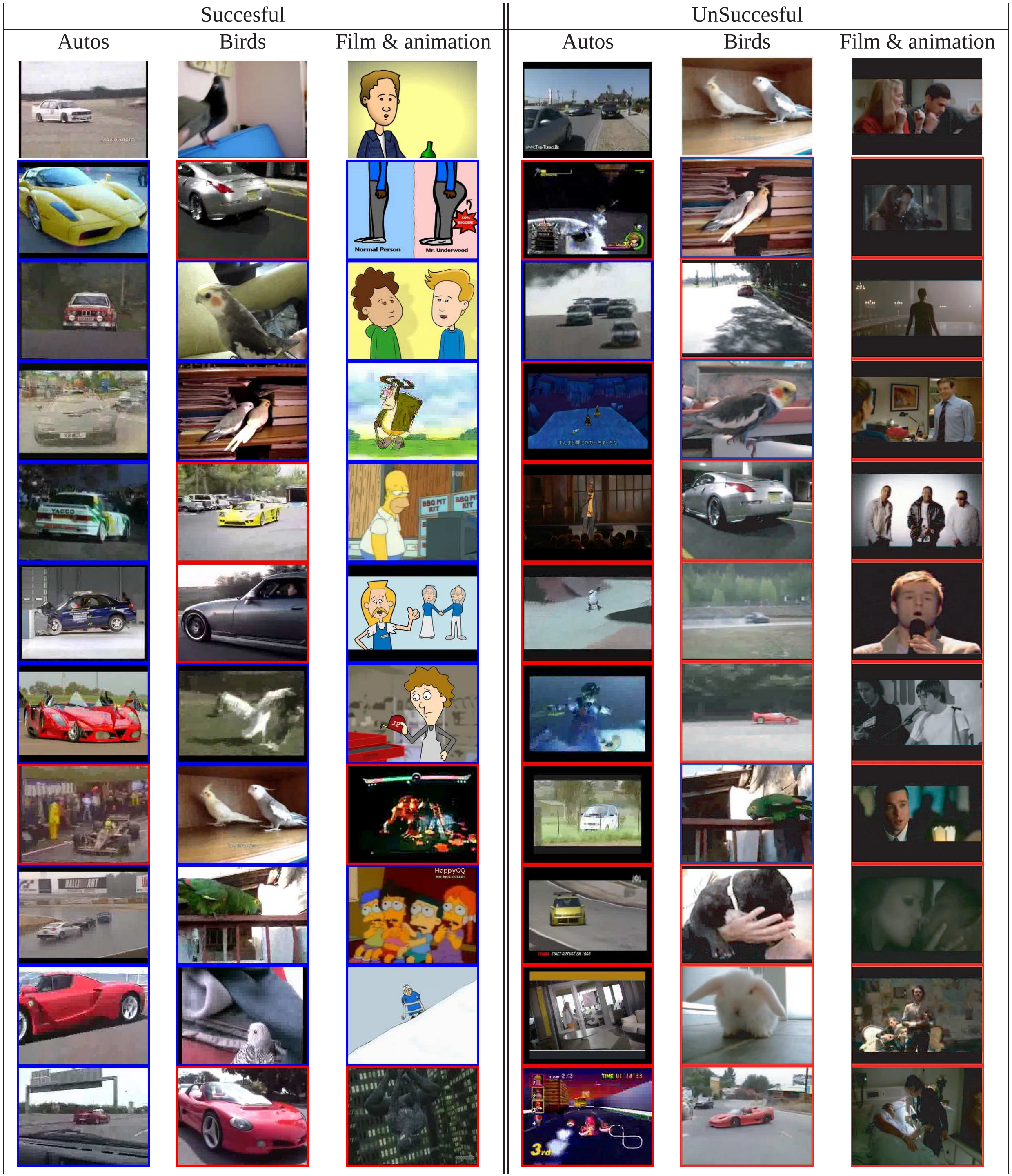

Figure 6. **Nearest neighbors.** Examples of nearest neighbor queries. The top image is the middle frame of a query video, the bottom 10 are from its nearest neighbors. Videos with the same (correct) category as the query video are bordered in blue and videos from different (wrong) category are bordered in red. Some videos that are marked wrong (red) seem to be correct candidates. This is due to the diversity within each category and the intersection between categories. For example, the nearest neighbors on the Film and Animation query on the last column come from the categories entertainment, music and Angelina Julie. Although highly similar in characteristics, these videos were tagged into different categories by their owners and are thus considered erroneous detections. Similarly, many of the videos in the sports category are of car races and thus show up as nearest neighbors of a clip from the Autos & Vehicles category.

[8] A. Ng, M. Jordan and Y. Weiss "On spectral clustering: Analysis and an algorithm" *In Advances in Neural Information Processing Systems 14*, 2001 6

[9] N. V. Patel and I. K. Sethi, "Video shot detection and characterization for video databases", *Pattern Recognition* Volume 30, Issue 4, , Image Databases, April 1997, Pages 583-592. 1

[10] Y. Rubner, C. Tomasi, and L. J. Guibas. "A Metric for Distributions with Applications to Image Databases" *Proceedings of the 1998 IEEE International Conference on Computer Vision*, Bombay, India, January 1998, pp. 59-66. 4

[11] A. F., Smeaton, P. Over and W. 2006. "Evaluation campaigns and TRECVid". *In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval* (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. 1

[12] S. M., Tahaghoghi, H. E., Williams, J. A., Thom, and T. Volkmer, 2005. "Video cut detection using frame windows". *In Proceedings of the Twenty-Eighth Australasian Conference on Computer Science* - Volume 38 (Newcastle, Australia). V. Estivill-Castro, Ed. ACM International Conference Proceeding Series, vol. 102. Australian Computer Society, Darlinghurst, Australia, 193-199. 1, 4

[13] A. Torralba and A. Oliva "Statistics of natural image categories" *Network: computation in neural systems*, Vol. 14, 391-412. 2003. 5

[14] A. Ulges, C. Schulze, D. Keysers, T. Breuel. "Content-based Video Tagging for Online Video Portals". *MUSCLE/ImageClef Workshop* 2007. 2, 3, 4, 5

[15] W. Xiong and J.C.M. Lee "Efficient Scene Change Detection and Camera Motion Annotation for Video Classification", *Computer Vision and Image Understanding*, Volume 71, Number 2, August 1998 , pp. 166-181(16) 1

[16] L. Zelnik-Manor and M.Irani "Statistical Analysis of Dynamic Actions", *IEEE Trans. on Pattern Analysis and Machine Intelligence* (PAMI), 28(9): 1530–1535, September 2006. 5, 6

[17] H. J. Zhang, J. W., Di Zhong and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing", *Pattern Recognition* Volume 30, Issue 4, , Image Databases, April 1997, Pages 643-658. 1

[18] W. Zhou, A. Vellaikal and C. C. Jay Kuo, "Rule-based video classification system for basketball video indexing", In *Proceedings of the 2000 ACM Workshops on Multimedia* (Los Angeles, California, United States), October 2000. 1