Kodak consumer video benchmark data set: concept definition and annotation

**Akira Yanagawa, *Alexander C. Loui, *Jiebo Luo, **Shih-Fu Chang, **Dan Ellis, **Wei Jiang, **Lyndon Kennedy, **Keansub Lee

*Research Laboratories
Eastman Kodak Company
Rochester, NY
{Alexander.loui, Jiebo.luo}@kodak.com

**Dept. Electrical Engineering,
Columbia University,
New York, NY
{akira, sfchang, dpwe, wjiang, lyndon, kslee}@ee.columbia.edu

Columbia University ADVENT Technical Report # 222-2008-8

September, 2008

This technical report includes the same content as the following paper, except information about data folders in Section 5 has been updated and the full list of concepts has been listed in the Appendix.

Alexander C. Loui, Jiebo Luo, Shih-Fu Chang, Dan Ellis, Wei Jiang, Lyndon Kennedy, Keansub Lee, Akira Yanagawa. Kodak's Consumer Video Benchmark Data Set: Concept Definition and Annotation. In ACM SIGMM International Workshop on Multimedia Information Retrieval, Germany, September 2007.

ABSTRACT

Semantic indexing of images and videos in the consumer domain has become a very important issue for both research and actual application. In this work we developed Kodak's consumer video benchmark data set, which includes (1) a significant number of videos from actual users, (2) a rich lexicon that accommodates consumers' needs, and (3) the annotation of a subset of concepts over the entire video data set. To the best of our knowledge, this is the first systematic work in the consumer domain aimed at the definition of a large lexicon, construction of a large benchmark data set, and annotation of videos in a rigorous fashion. Such effort will have significant impact by providing a sound foundation for developing and evaluating large-scale learning-based semantic indexing/annotation techniques in the consumer domain.

This report includes information about the concept definitions, the annotation process, video collection process, and the data structures used in the release file. The released dataset includes the annotations, extracted visual features (for videos from Kodak), and URLs of videos from YouTube. The Appendix section also includes the full list of concepts (more than 100 concepts in 7 categories) that have been defined in the consumer video domain.

Categories and Subject Descriptors

Information Storage and Retrieval – Collection, Standards; Database Management – multimedia databases, image databases

General Terms

Standardization, Management, Human Factors, Measurement

Keywords

Video classification, semantic indexing, consumer video indexing, multimedia ontology

1. INTRODUCTION

With the prevalent use of online search engines, most users are now accustomed to simple and intuitive interfaces when interacting with large information sources. For text documents, such simple interfaces may be handled by the typical keyword search paradigm. However, for other domains that involve multimedia information, novel techniques are required to index content at the semantic level, addressing the well-known problem of semantic gap. The need for semantic-level indexing is especially obvious for domains such as consumer video because of the lack of associated textual metadata and the difficulty of obtaining adequate annotations from users. To solve this problem, one emerging research area of semantic indexing is the development of automatic classifiers for annotating videos with a large number of predefined concepts that are useful for specific domains. To provide a sound foundation for developing and evaluating large-scale learning-based semantic indexing/annotation techniques, it is important to apply systematic procedures to establish large-scale concept lexicons and annotated benchmark video data sets. Recently, significant developments for such purposes have been made in several domains. For example, NIST TRECVID [1], now in its sixth year of evaluation, has provided an extensive set of evaluation video data set in the broadcast news domain. It includes hundreds of hours of videos from multilingual broadcast news channels. To define a common set of concepts for evaluation, a recent effort has also been completed to define a Large-Scale Concept Ontology for Multimedia (LSCOM) [2], which includes 834 concepts jointly selected by news analysts, librarians, and researchers. A subset of these concepts (449) has been annotated through an exhaustive manual process over the entire 2006 TRECVID development set [3]. Availability of such a large-scale ontology and fully annotated video benchmark data set has proved to be very valuable for researchers and system developers. So far, about 200 research groups have downloaded the LSCOM definition and annotation set. In addition, large-scale baseline automatic classifiers for LSCOM concepts, such as Columbia374 [4] and MediaMill 491 [5], have been developed and broadly disseminated in the research community.

Significant efforts have also been made in other domains to establish large-scale benchmark data sets for image search and object recognition. For example, Caltech 101 [6] includes 101 categories of images downloaded from the Web to evaluate performance of object recognition techniques. ImageCLEF [7] includes a large set of medical images and web images for evaluating image retrieval methods.

However, for consumer videos, to the best of our knowledge, there has been no systematic effort so far to develop large-scale concept lexicons and benchmark data sets. Although automatic consumer video classification has been reported in the literature, most of the prior work dealt with few concepts and limited data sets only. To contribute to the research of consumer video indexing and annotation, we have developed Kodak's consumer video benchmark data set, including a significant number of videos (a few thousand) from actual users who participated in an extensive user study over a one-year period and from a user-generated content site (YouTube). It also includes a lexicon with more than 100 semantic concepts and the annotations of a subset of concepts over the entire video data set. The concepts have been chosen in a systematic manner, considering various criteria discussed below. As far as we know, this is the first systematic work in the consumer domain aimed at the definition of a large lexicon, construction of a large benchmark data set, and annotation of videos in a rigorous fashion.

It is nontrivial to determine the appropriate lexicon of semantic concepts for consumer videos, as the correct lexicon may depend highly on the application. To fulfill the needs of actual users, we adopt a user-centric principle in designing the lexicon. The concepts are chosen based on findings from user studies confirming the usefulness of each concept. In addition, we consider the feasibility of automatic detection and concept observability in manual annotation. Our lexicon is broad and multi-categorical, including concepts related to activity, occasion, people, object, scene, sound, and camera operation. It also includes concepts manifested by multimodal information. That is, our concepts may be visual-oriented and/or audio-oriented. To ensure the quality of annotation, we adopt a multi-tier annotation strategy for different classes of concepts. Some concepts use keyframe-based approaches to maximize the annotation throughput. For others, playback of an entire video clip is required to judge the presence of the concepts.

In this paper we will describe details of Kodak's consumer video benchmark data set. In Section 2, we introduce the principles for designing the lexicon and the definitions of the selected concepts; Section 3 describes the video data set and the procedures for extracting keyframes; Section 4 presents the manual procedures for concept annotation and some results of annotation quality analysis; Section 5 includes information about the data structure and file system of the released data set; and in Section 6 we conclude our work and give some further discussion.

2. LEXICON AND CONCEPTS

The lexicon used in Kodak's consumer video benchmark data set was constructed based on an ontology derived from a user study conducted by Eastman Kodak Company. The ontology consists of 7 categories: SUBJECT ACTIVITY, ORIENTATION, LOCATION, TRADITIONAL SUBJECT MATTER, OCCASION, AUDIO, CAMERA MOTION. Under these categories, over 100 concepts are defined based on feedback from user studies confirming the usefulness of each concept. An example of the categories and concepts is shown in Table 8 in the Appendix. The full list of categories and concepts can be found in Appendix. This ontology has been chosen through three steps. First, an earlier user study based on a large collection of consumer photos has been conducted by Kodak to discover concepts interesting to users in practical applications. These concepts are used to form the initial candidate lexicon for the consumer video data set. Second, the initial concept list was refined based on a smaller-scale user study to find interesting concepts for consumer videos. A relatively smaller collection of video data, compared to the photo collection, was used. Finally, the availability of each selected concept (the number of videos we may obtain from users for each concept) is investigated, and the rare concepts are excluded.

Because of the limitation of both the annotation and the computation resources, in this version 25 concepts are further selected from Kodak's ontology based on 3 main criteria: (1) visual and/or audio detectability—whether the concept is likely to be detected based on the visual and/or audio features; (2) usefulness—whether the concept is useful in practical consumer media application; (3) observability—whether the concept is observable by the third-person human annotators through viewing the audio-video data only. Such criteria are identical to those used in selecting the large-scale semantic concepts for broadcast news in LSCOM [2]. In addition, we also consider one additional criterion, availability, i.e., the number of video clips we may expect to acquire for a concept from actual users. To estimate the availability of a concept, we conduct searches using concept names as keywords against YouTube and AltaVista, two popular sites for sharing user-generated videos. The number of the retuned video clips in the search results is used to approximate the availability of a concept.

The final lexicon used in Kodak's consumer video benchmark data set contains 25 concepts as shown in Table 1. Note these concepts are multimodal in nature—some are primarily manifested by the visual aspect (e.g., night, sunset), some are audio-oriented (e.g., music, singing), and others involve both visual and audio information (e.g., wedding and dancing).

Table 1: Selected concepts and definitions

	Concept	*Definition*
activities	dancing	One or more people dancing
	singing	One or more people singing. Singer(s) both visible and audible. Solo or accompanied, amateur or professional.
occasions	wedding	Videos of the bride and groom, cake, decorated cars, reception, bridal party, or anything relating to the day of the wedding.
	birthday	This event is typically portrayed with a birthday cake, balloons, wrapped presents, and birthday caps. Usually with the famous song.
	graduation	Caps and gowns visible
	ski	Emphasize people in action (vs. standing)
	picnic	Video taken outdoors, with or without a picnic table, with or without a shelter, people, and food in view.
	show	Concerts, recitals, plays, and other events.
	parade	Processing of people or vehicles moving through a public place
	sports	Focus initially on the big three: soccer, baseball/softball, and football
	playground	Swings, slides, etc. in view
	park	Some greenery in view
	museum	Video is taken indoors and is of exhibitions of arts, crafts, antiques, etc.
scene	sunset	The sun needs to be in front of the camera (although not necessarily in view)
	beach	Largely made up (1/3 of the frame or more) of a sandy beach and some body of water (e.g., ocean, lake, river, pond). Note "beach" should be explicitly called out. In a more strict definition, a "beach" scene contains at least 10% each of water, sand, and sky, and was taken from land. Pictures taken primarily of water from a boat should be called "open water".
	night	The video is taken outdoors at night (after sunset).
object	people -- 1	One person: the primary subject includes only one person.
	people -- 2	Group of two: the primary subject includes two people.
	people -- 3	Group of three or more: the primary subject includes three or more people. This description applies to the primary subject and not to incidental people in the background.

	animal	Pets (e.g., dogs, cats, horses, fish, birds, hamsters), wild animals, zoos, and animal shows. Animals are generally "live" animals. Those stuffed or mounted (taxidermy) may qualify depending on how "lively" they look.
	boat	Boat in the water
people	crowd	The primary subject includes a large number of people in the distance.
	baby	Infant, approximately 12 months or younger
sound	music	Clearly audible professional or quality amateur music in the soundtrack (which may also include vocals and other instruments). There is emphasis on the quality of the music.
	cheer	One or more people cheering - shouts of approval, encouragement, or congratulation.

3. VIDEO DATA SETS AND KEYFRAMES

Kodak's consumer video benchmark data set includes two video subsets from two different sources. Kodak's video data set includes 1358 consumer video clips contributed by users who participated in the user study; and the YouTube video data set includes consumer video clips downloaded from the YouTube website. In the following subsections, we will describe both data sets in detail.

3.1 Kodak's Video Data Set

Kodak's video data set was donated by actual users to Eastman Kodak Company for research purposes. The vast majority of the videos were recorded by either the Kodak EasyShare C360 zoom digital camera or the Kodak EasyShare V570 dual lens digital camera. The videos were collected over the period of one year from about 100 users, thus spanning all seasons and a wide variety of occasions. It is also geographically diverse as the majority of users took videos outdoors and away from home, including trips across the US and also overseas. These users were volunteers who participated in typically three-week-long camera handouts. They represent different consumer groups (e.g., proactive sharers, conservative sharers, prints-for-memory makers, digital enthusiasts, and just-the-basics users) and were identified through a web-based survey. Female users slightly outnumbered male users. A unified video format, MPEG-1, is used to allow easy handling of videos. The videos whose original format is QuickTime movie or AVI format were transcoded to MPEG-1 format according to original bit rates. Other detailed information about this data set is shown in Table 2. More details about the data structure and file formats will be introduced in Section 5.

Table 2: Information of Kodak's video data set

Total Number of Video Clips	1358	
Total Number of Key Frames	5166	
Lengths of Videos	Min	0.1 s
	Max	393.1 s
	Avg	31.1 s
Resolution	640 × 480 or 320 × 240 (pixels)	
Video Format	MPEG-1	
Bit Rates (Audio + Visual)	Min	280 kb/s
	Max	19,115 kb/s
	Avg	7.99 kb/s
Frame Rate	30 frames/s	
Audio Sampling Rate	44100 Hz	

3.2 YouTube Video Data Set

The YouTube video data set was downloaded by searching over the YouTube online system with keywords derived from the concept names. For some concepts we directly use the concept name as the search keyword. But for other concepts we need to modify the concept names or add additional words in order to retrieve videos of the intended semantics. For example, when we used "cheer" as a keyword to find videos for the "cheer" concept, YouTube returned many videos of cheerleaders. In such a case, we expanded the search keywords to include "cheer, cheer up" based on the concept definition and subjective interpretation of the concept in order to increase the chance of retrieving videos relevant to the concept. In Table 3, the actual keywords used for searching videos on YouTube are listed. Then from the result list returned by the YouTube search engine, the top most relevant videos were downloaded and then further screened manually to ensure their relevance to the concept. The final number of videos stored for each concept are also listed in Table 3.

As with Kodak's video data set, the downloaded video clips were transcoded to 200 Kbps MPEG-1 format with the frame rate 30 fps. Other detailed information of this data set is described in Table 4. In addition, unlike Kodak's videos, an additional file is provided for each video clip to record the relevant metadata information, including the URL link of the video and thumbnail image, the name of the author(s), the tags, the title, and the category. An example of the image and metadata is given in Figure 1. Note that we did not extract keyframes for YouTube video data, and the final annotation for each concept is associated with each video clip rather than with individual keyframes.

Table 3: Keywords and number of videos from YouTube

	Concept	*Keywords*	*# of videos downloaded*	*After manually filtering*
activities	dancing	Dancing	189	101
	singing	Singing	192	95
occasions	wedding	Wedding	196	86
	birthday	Birthday	192	101
	graduation	Graduation and caps and gowns	191	107
	ski	Ski	195	77
	picnic	Picnic	187	97

	show	Show, Concert, Play, Event	196	54
	parade	Parade	194	113
	sports	Soccer, Basketball Football, Baseball, Volleyball, Ping-pong	340	95
	playground	Playground	194	80
	park	Park	191	74
	museum	Museum	192	63
scene	sunset	Sunset	179	72
	beach	Beach	183	105
	night	Night	193	79
object	people	People	187	48
	animal	Pets, Animal	198	31
	boat	Boat	191	98
people	crowd	Crowd	191	71
	baby	Baby	184	81
sound	music	Music	197	59
	cheer	Cheer, Cheer up	187	86

3.3 Keyframe Sampling (Kodak's Data Set)

From the videos in Kodak's video data set, we sample keyframes based on a uniform time interval, i.e., 1 keyframe per 10 s. Based on the experience obtained from the user study, we consider the 10 s sampling interval to be a good tradeoff between computation/storage requirements and indexing accuracy. For static concepts (e.g., locations and occasions), we assume that the

http://www.youtube.com/get_video?video_id=MQ1ceUTzbFA&l=38&t=OEgsToPDskKXdiSHaPWqTYv25IwOSgwT
http://www.youtube.com/watch?v=MQ1ceUTzbFA
http://sjl-static13.sjl.youtube.com/vi/MQ1ceUTzbFA/2.jpg
sunset in Hawaii
hawaii sunset
Travel & Places
Kellyraaz

Figure 1: An image and metadata example for YouTube data

Table 4: Additional information of YouTube video data set

The Number of Video Clips	1874	
The Length of the Videos	Min	0.1 s
	Max	2573.7 s
	Avg	145.1 s
Resolution	320 × 240	
Video Format	MPEG-1	
Bit Rates (Audio+Visual)	200 Kbps	
Frame Rate	30 frames/s	
Audio Sampling Rate	44100 Hz	

video content will not change much in each 10 s interval. In such a case, keyframes will be sufficient for analyzing the concept. However, for other concepts (e.g., dancing), information in the temporal dimensions (object movements and dynamics) needs to be considered. In this case, features need to be extracted from image frames at a much higher rate. In other words, the above-mentioned coarsely sampled keyframes are intended for analyzing the static concepts only. Concepts involving strong temporal information need to be analyzed using the video clips. Note for audio-based analysis, typically the sound track of the entire video clip is used, rather than just audio signals associated with the keyframes. However, in practice we may extract audio signals just near the keyframe time point in order to combine the local audio cues with the visual cues found near the keyframe time point.

To ensure quality of the extracted keyframes, we deliberately insert an initial time offset to the sample schedule. An offset of 1 s is applied at the beginning because the first frames (time = 0) of some video clips are totally black or blurred. In other words, keyframes are extracted at the following time points: 1 s, 11 s, 21 s, 31 s, etc. In addition, to avoid missing important content, if the duration of a video clip is less than 11 s, the final frame of the clip will be included automatically.

Here is a summary of the keyframe sampling procedure.

D (s) is the duration of a video clip:

a) $D < 1$: 1 keyframe is extracted, namely the last frame.

b) $1 \leq D < 11$: two keyframes are extracted. One at 1 s and the other at the last frame.

c) $D > 11$: keyframes are extracted at time points = 1 s, 11 s, 21 s, 31 s, etc.

Although we could have used an automatic keyframe-extracting algorithm, we did not do so because the algorithm has not been fully evaluated and does not always produce consistent results. Using the simple temporal subsampling technique described above at least ensures consistency.

4. ANNOTATION

In this section, we will describe the details on how we obtain the ground-truth annotation for Kodak's video data set and the YouTube video data set, respectively.

4.1 Annotation for Kodak's Video Data Set

The concept labels for Kodak's video data set are manually annotated by students at Columbia University. To increase the throughput of the annotation process and ensure good quality of the resulting labels, we employed a multi-tier annotation strategy. For visual-oriented concepts (e.g., activities, occasions, scenes, people, objects), we always obtain annotations of individual keyframes using keyframe-based annotation tools. Such an approach is sufficient for most static concepts (see Table 5). For concepts that involve information in the temporal dimension, we further employ video playback tools to verify the correctness of the label. We do this in an incremental manner; namely, only

those keyframes receiving positive labels in the first step are included in the video-based verification process. During the verification step, an annotator plays back each candidate video clip and marks the presence or absence of the concept. Keyframes corresponding to negative videos (where the concept is absence) are corrected as negative. In this case, it is possible that only a subset of keyframes of a video receive positive labels, while the remainder are negative. We use the above incremental procedure to avoid the high workload involved in using video playback to annotate every clip. Based on our experience, the keyframe-based annotation process is much faster than the video-based process. On average, the throughput of the keyframe-based annotation process is about 1–3 s per keyframe, while the throughput for the video-based annotation is about 30–60 s per video. Finally, for audio-oriented concepts (e.g., music and cheer), we use the video-based annotation process to label every video clip. Binary labels are assigned for each concept—presence or absence.

An alternative approach to annotating concepts in videos is to play back the video and audio tracks and mark the boundary information of the concept. There are several well-known tools available in the literature for such a purpose, but they are usually time-consuming. In this version of data set, we decide to adopt the above multi-tier labeling process and review the need for finer granular labels in the future.

The annotation strategies used for different concepts are shown in Table 5. Table 6 and Figure 2 show the number of positive and negative keyframes and videos for each concept in ground-truth annotation.

To annotate Kodak's video data set, we utilized two tools. The first one is for annotation based on only keyframes (Figure 3). This tool is developed by the CMU Informedia group [8]. In this annotation tool, multiple keyframes are shown at the same time, and an annotator judges whether a specific concept is present in each keyframe. The annotator may enter labels individually for each keyframe on the screen either by clicking with the mouse or using keyboard shortcuts. The second tool is for annotation based on video playback (Figure 4). This tool shows a video clip and an annotator can repeat, pause, skip, and stop the video using the tool. The annotator goes through each video clip one-by-one.

Table 5: Annotation strategies

	Concept	*Annotation Strategy*
activities	dancing	Keyframes + Video
	singing	Video
occasions	wedding	Keyframes
	birthday	Keyframes
	graduation	Keyframes
	ski	Keyframes + Video
	picnic	Keyframes
	show	Keyframes
	parade	Keyframes + Video
	sports	Keyframes
	playground	Keyframes
	park	Keyframes
	museum	Keyframes
scene	sunset	Keyframes
	beach	Keyframes
	night	Keyframes
object	one person	Keyframes
	group of two	Keyframes
	group of three or more	Keyframes
	animal	Keyframes
	boat	Keyframes
people	crowd	Keyframes
	baby	Keyframes
sound	music	Video
	cheer	Video

Table 6: The number of positive and negative keyframes and video clips on Kodak's video data set

Concept	*# Positive Keyframes*	*# Negative Keyframes*	*# Positive Videos*	*# Negative Videos*
animal	186	4980	69	1289
baby	140	5026	38	1320
beach	74	5092	37	1321
birthday	54	5112	15	1343
boat	96	5070	39	1319
crowd	448	4718	144	1214
dancing	226	4940	48	1310
graduation	15	5151	3	1355
group of 3+	689	4477	246	1112
group of two	437	4729	171	1187
museum	52	5114	18	1340
night	240	4926	87	1271
one person	1054	4112	374	984
parade	103	5063	25	1333
park	407	4759	150	1208
picnic	22	5144	13	1345
playground	78	5088	24	1334
show	321	4845	54	1304
singing	99	5067	50	1308
ski	433	4733	151	1207
sports	54	5112	21	1337
sunset	141	5025	27	1331
wedding	186	4980	69	1289
cheer	N/A	N/A	175	1183

music	N/A	N/A	206	1152

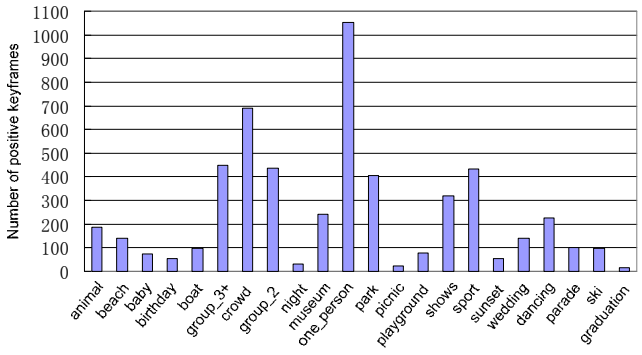


Figure 2: Numbers of positive keyframes for Kodak's video data set. Concepts with video only annotation are not included

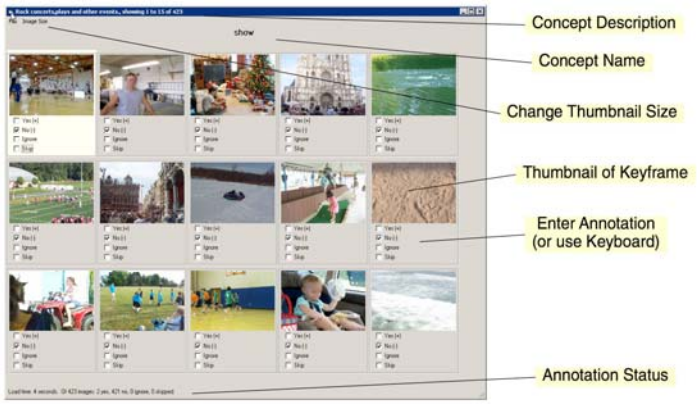


Figure 3: Example of the annotation tool from the CMU Informedia group for keyframe annotation

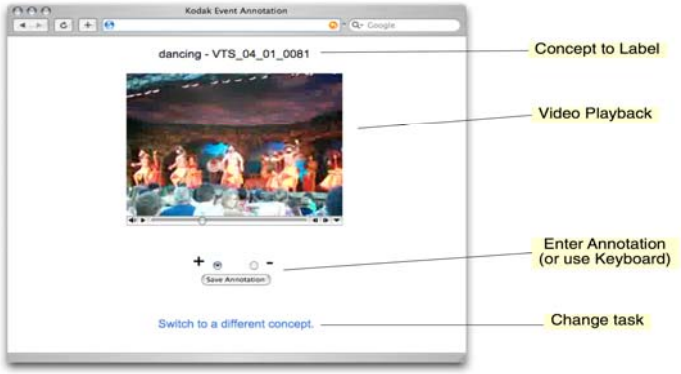


Figure 4: Annotation tool for video clips

4.2 Annotation for YouTube Video Data Set

For the 4539 videos (about 200 videos per concept) downloaded from the YouTube website, we first manually pruned out the commercial videos (which are different from our focus on consumer videos) and the low-quality videos (especially those having poor sound quality). After pruning, only 1873 (41%) video clips remained. Then we annotated these 1873 video clips according to the 25 concepts described earlier (as defined in Section 2) at the video level by viewing the entire video clips. This method tends to assign more concepts per video than the keyframe-based annotation method, because all the frames were taken into account and the chance of finding a concept in some part of the video increases (if any part of the video contains a concept, the whole video clip is considered as containing this concept). Table 7 and Figure 5 list the number of positive and negative video clips for every concept.

Table 7: Numbers of positive and negative video clips for each concept over the YouTube video data set

Concept	*# Positive Videos*	*# Negative Videos*
animal	61	1812
baby	112	1761
beach	130	1743
birthday	68	1805
boat	89	1784
crowd	533	1340
dancing	189	1684
graduation	72	1801
group of 3+	1126	747
group of two	252	1621
museum	45	1828
night	300	1573
one person	316	1557
parade	91	1782
park	118	1756
picnic	54	1819
playground	96	1777
show	211	1662
singing	345	1529
ski	68	1805
sports	84	1789
sunset	68	1805
wedding	57	1816
cheer	574	1299
music	653	1220

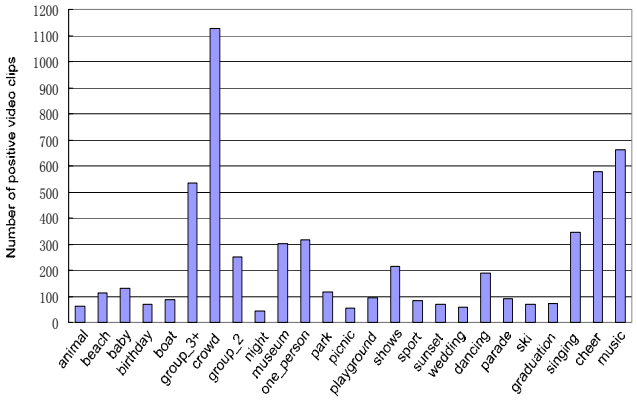


Figure 5: The number of positive samples on YouTube video data set (video clips)

4.3 Assessment of Consistence in Annotation (Kodak's Data Set)

Different annotators (observers) may have different judgments for some concepts during the annotation process, and this may cause inconsistency of the annotations. There are several possible reasons why different annotators have different opinions. First, the interpretations of some concepts are actually quite subjective and dependent on the annotator's knowledge. For example, for the "baby" concept, it is sometimes difficult to determine whether a child shown in a video satisfies the definition (i.e., child less than 1 year old) from only the visual appearance and different people have different opinions. Second, annotation based on keyframe only, although typically adequate for some concepts, is insufficient in some cases. For example, based on the consideration of throughput, we used keyframe-based annotation for the concept "wedding". But this indeed has caused ambiguity and resulted in different labels from different users in some cases. Third, human annotation is not error free and mislabeling indeed occurs.

Therefore, it is important to investigate the relationships between different observers' annotations. One specific way is to measure the degree of agreement among labels from different users. This will help to assess the quality of the annotations, which are affected by many factors discussed above.

In this subsection, we analyzed the inter-annotator agreement for keyframe-based annotations over 19 concepts in Kodak's data set. To do this, we have arranged that each concept was annotated by 2 annotators for 20% of Kodak's data set—one person annotated the entire set and the other one annotated an overlapped subset that consisted of 20% of videos randomly selected from the entire set. Then Kappa coefficient [9] was used to measure the consistency among the observers while excluding the probability of consistency by chance. The larger the Kappa value is, the better the consistency is among different annotators. Specifically, Kappa value is defined by the following equation:

$$\boldsymbol{Kappa} = \frac{\Pr(\boldsymbol{I}) - \Pr(\boldsymbol{C})}{1 - \Pr(\boldsymbol{C})},$$

where Pr(I) is the probability of the agreement among observers and Pr(C) is the probability of the coincidence by chance. In this analysis, we set the prior probability (i.e., chance of finding positive labels) of a concept to be its Pr(C).

The Kappa values for different concepts are showed in Figure 6. Kappa values greater than 0.6 usually are considered to be good [10]. From the results, the Kappa values of "crowd," "playground," "wedding," "birthday," and "picnic" are less than 0.5. This may be caused by the ambiguity of the concept definitions. For example, different people interpreted the concept "crowd" differently partly because the definition of "crowd" does not specify how many people comprise a crowd. Annotations of the "playground" concept may vary depending on the interpretation of the requirements of having certain structures or objects in view. Also, as mentioned earlier, some concepts such as "wedding" and "birthday" may suffer from using keyframes only in the annotation process. On the other hand, some concepts such as "one person," "sports," "show," "night," "boat," and "museum," have very good inter-subject agreement with the Kappa values over 0.7. Such results are intuitive and reasonable because these concepts have clearer and simpler definitions than those with low agreement.

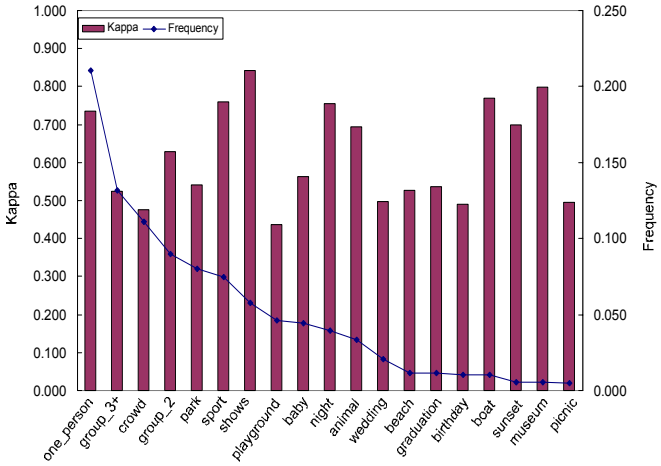


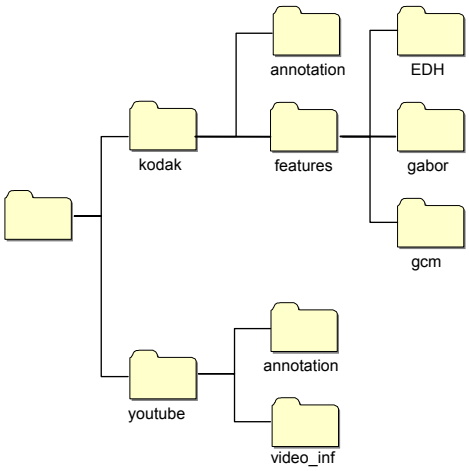
Figure 6: Kappa values

It is also interesting to observe that Kappa values do not correlate with concept frequency. This is consistent with the findings reported in our prior project on annotation of concepts for news videos [2]. Our conclusion is that inter-subject annotation consistency depends mostly on the clarity (unambiguity) of the definitions and the effectiveness of the annotation tool.

5. DATA STRUCTURE

In this section, we will introduce the data structure for organizing both metadata and ground-truth annotations. Figure 7 shows the folder structure. Under the root named "consumervideo" folder, there are two folders: "kodak" and "youtube." Each folder contains subfolders where annotations and the visual features and the information of the YouTube's video are stored. In the following subsections, we will describe the data structure for each subfolder respectively.

Figure 7: Data Structure of the benchmark data set



5.1 Video Data Folder

For YouTube video data, to reuse these videos conveniently, an information file is provided for each video clip. The name of the file is the same as the name of the video file, but with extension "vinf." This file includes additional information downloaded from YouTube, such as the URL link of the video and thumbnail, the name of author(s), the tags, the title, and the category. The format of the file is described below.

Line 1: [Don't care: Internal Use Only]

Line 2: [URL of the video]

Line 3: [Don't care: Internal Use Only]

Line 4: [Title]

Line 5: [Tags] pets dogs animals

Line 6: [Category]

Line 7: [Author]

5.2 Annotation Data Folder

The annotation data is placed under the "annotation" subfolder, in both the Kodak folder and the YouTube folder. The annotation file is a simple text file. The name of the annotation file is described below.

[xxx].txt

xxx: the name of concept, e.g. one_person, night, and so on.

The format of the annotation file is described below.

Annotation(1 or -1) [TAB]%[TAB]**Keyframe File name**

Annotation(1 or -1) [TAB]%[TAB]**Keyframe File name**

Annotation(1 or -1) [TAB]%[TAB]**Keyframe File name**

Annotation(1 or -1) [TAB]%[TAB]**Keyframe File name**

If the value of annotation is "1," the video is positive, i.e., relevant to this concept; and if the value of annotation is "-1," the video is negative, i.e., irrelevant to this concept.

5.3 Visual Features (Only for Kodak's Data)

Three visual features are placed in "features" subfolders: edged direction histogram (EDH), Gabor (GBR), and grid color moment (GCM). For information about these features in detail, please refer to [11].

The "features" folder contains data files with the features that we have used. Under "features" folder, there are three folders named "EDH", "gabor", and "gcm," which contain data files for the edge direction histogram (EDH), Gabor texture (GBR), and grid color moment (GCM) features, respectively, as shown in Figure 7. The file names for the visual features are as "kodak.edh," "kodak.gbr "and "kodak.gcm."

In the file, feature vectors are sorted row-wise by keyframe, and are sorted column-wise by element. This format is described below.

The element of the features →

Element1[tab]**Element2**[tab]...[tab]%**Keyframe File name**

Element1[tab]**Element2**[tab]...[tab]%**Keyframe File name**

Element1[tab]**Element2**[tab]...[tab]%**Keyframe File name**

Element1[tab]**Element2**[tab]...[tab]%**Keyframe File name**

The number of the elements of EDH, GBR, and GCM are 73, 48 and 225 respectively [11].

6. CONCLUSION AND FUTURE WORK

In this paper we presented an actual consumer video benchmark data set that includes a rich consumer-based lexicon and the annotation of a subset of concepts over the entire video data set. This is a first systematic work in the consumer domain that aims at the definition of a large lexicon, construction of a large benchmark data set, and annotation of videos in a rigorous fashion. This effort will provide a sound foundation for developing and evaluating large-scale semantic indexing/annotation techniques in the consumer domain. A preliminary evaluation of semantic classifiers using this large data set is described in another paper of this special session. We plan to expand the lexicon by considering outcomes of consumer-based user studies and to discover related concepts from online user-contributed sites.

7. ACKNOWLEDGMENTS

Funding for the annotation process has been provided by Eastman Kodak Company. The annotation work of Kodak's data set was completed by Columbia University students Nelson Wei and Jing Jin, under the coordination of Lyndon Kennedy. Annotation of audio concepts over Kodak's data set and all concepts over the YouTube data set was accomplished by Keansub Lee. Steve Sitter and Deniz Schildkraut of Kodak helped with both the data collection and concept detection. We thank CMU Informedia group for sharing their keyframe annotation tool.

8. REFERENCES

[1] NIST. TREC video retrieval evaluation (TRECVID). 2001-2006, http://www-nlpir.nist.gov/projects/trecvid/.

[2] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. IEEE MultiMedia, vol. 13 (2006), pp. 86–91.

[3] LSCOM Lexicon Definitions and Annotations Version 1.0, Columbia University ADVENT Technical Report #217-2006-3, March 2006. (http://www.ee.columbia.edu/dvmm/lscom)

[4] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Columbia University ADVENT Technical Report # 222-2006-8, March 2007. http://www.ee.columbia.edu/dvmm/columbia374.

[5] M. Worring, C. Snoek, O. de Rooij, G.P. Nguyen, and A. Smeulders. The MediaMill Semantic Video Search Engine. IEEE ICASSP, (April 2007), Hawaii.

[6] Caltech 101 data sets, http://www.vision.caltech.edu/Image_Datasets/Caltech101

[7] The CLEF Cross Language Image Retrieval Track (ImageCLEF), http://ir.shef.ac.uk/imageclef/ .

[8] The Informedia Digital Library Project. http://www.informedia.cs.cmu.edu.

[9] J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, vol. 20, no. 1 (1960), 37–46.

[10] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. Biometrics. vol. 33, no, 1 (1977), 159–174.

[11] A. Yanagawa, W. Hsu, and S.-F. Chang, "Brief Descriptions of Visual Features for Baseline TRECVID Concept Detectors," Columbia University ADVENT Technical Report #219-2006-5, July 2006.

9. APPENDIX

The following table shows the complete video concept ontology for the consumer domain. It includes 155 concepts from 7 categories, among which 25 initial concepts have been selected in annotation and video data collection.

Table 8: The complete list of concepts and definitions of Kodak Consumer Video Concept Ontology

	Concept	*Definition*
SUBJECT ACTIVITY	Posed	The person intentionally assumes or holds a particular position or posture while being videoed. If the person(s) realized their video is being taken, they are posed. A child less than 12 months old or an animal is recorded as posed only when he or it is deliberately being held or positioned for the purpose of being videoed.
	Candid	Still -The subject is not posing and appears without movement. Inanimate objects, such as buildings and structures, are included in this category.
	Candid Motion	The subject is not posing, but was videoed while in action, in the process of a change in position, or while moving a part of the body. It also includes moving cars, trains, etc.
	Dancing	
	Eating	a more strict concept of "Dining" involves multiple people sitting at a dining table with plates and/or food (home or restaurant)
	Playing	
	Riding	
	Running	
	Singing	
	Sitting	
	Sleeping	
	Sports	focusing initially on big three: soccer, baseball/softball, and football
ORIENTATION	Horizontal	Videos are of the bride and groom, cake, decorated cars, reception, bridal party, or anything relating to the day of the wedding.
	Vertical Right Up	This event is typically portrayed with a birthday cake, balloons, wrapped presents, and birthday caps. Usually with the famous song.
	Vertical Left Up	caps and gowns
	Upside Down/Deliberate Tilt	emphasizing people in action (vs. standing)
	Horizontal to Vertical Shift	The video is taken outdoors, with or without a picnic table, indoors (shelter/pavilion) or outdoors (probably should separate the two), people and food in view.

	Vertical to Horizontal Shift	Rock concerts, plays and other events.
LOCATION	Outdoors	The video is taken outdoors during the day.
	Outdoors Night	The video is taken outdoors at night (after sunset).
	Indoors-to-Out	The video is taken of a subject outdoors from inside a house or shelter, sometimes through a window. While the camera is indoors, the outdoor lighting in this case will most affect the video. If through a window, "Subject Through Glass" should be recorded for that frame in the Subject Matter section along with whatever the subject is.
	Living/Family Room	The video is taken indoors in a room which may contain items such as a television, rug, floor, couch and related living room furniture, fireplace, piano, etc.
	Kitchen	The video is taken indoors in a room which may contain items such as a range or other kitchen appliances, sink, counter, and wall cabinets. A dinette not separated from a kitchen by some permanent wall structure falls under this category.
	Dining Room	The video is taken indoors in a room which may contain items such as a table, china cabinet, hutch, etc.
	Bedroom	The video is taken indoors in a room which may contain items such as a bed, pillows, and other related bedroom furniture.
	Other Indoors-Home	The video is taken indoors in any other room in the home that does not fit into one of the above categories.
	Party House	The video is taken indoors in a large area which is seating a group of people participating in some social activity, such as a wedding or banquet.
	Theater/Auditorium	The video is taken indoors and may contain items such as a stage, tiers of seats, a concert, a play, a motion video, etc.
	Church (Sanctuary)	The video is taken indoors and may contain items such as an altar, a pulpit, rows of benches, ornate walls and ceilings, stained glass, statues, etc. The video is taken within a church, but not including a church meeting room or office.
	Business Office/Industry	The video is indoors and may contain items such as a desk, machinery, bookcases, etc.
	Retailer/Restaurant. Etc	The video is taken indoors in a department store, mall, specialty shop, location with a number of separate dining tables set up for small groups, etc.
	Museum	The video is taken indoors and is of an exhibition of arts, crafts, antiques, etc.
	School Room	The video is taken indoors and may contain such items as chairs, student desks, large windows decorated with classroom paraphernalia, school cafeteria, etc.
	School Fieldhouse	The video is taken indoors in a gymnasium, gym locker room, indoor swimming pool, etc.
	Professional Sports Arena	The video is taken in an indoor stadium where a professional hockey game, basketball game, rodeo, entertainment spectacles (e.g., as a circus or "Ice Follies") are held.
	Picnic Pavilion/Shelter	The video is taken indoors in a gazebo, park shelter, party tent, or covered porch of a house.
	Theme Park Buildings	The video is taken indoors in a building used for displays or exhibits, such as Disney World's "It's A Small World"

		building. Also includes videos inside any theme park (e.g., Busch Gardens, Six Flags Over Texas) buildings.
	Restoration Buildings	The video is taken indoors in an old structure restored to original state (e.g., "Pilgrim Village," "Genesee Country Village," "House of Seven Gables").
	Hotel/Dorm Room	The video is taken indoors in a hotel or dorm room. The distinguishing characteristic of a hotel/motel room is a somewhat compact array of varied function furniture (table, bed, television, or couch).
	Hospital	The video is taken indoors in the baby's nursery, a hospital room, etc.
	Airport Terminal	The video is taken indoors and may contain long corridors, large windows, airline insignias, etc. This category may also include videos taken inside a bus terminal and a train terminal.
	Other Public Buildings	The video is taken indoors at a public building which cannot be classified into one of the above categories, such as at a bowling alley, zoo, aircraft hangar, courthouse, post office, church basement, hotel hallway, hotel lobby, roller skating rink, ice skating rink, etc.
	In to Out	A video of a scene outside a stationery or moving vehicle taken from a covered or an enclosed compartment of that vehicle. An example is a video of a landscape taken from a vehicle, typically through a window with glass. The frame of the window may be visible. If the subject is taken through glass, record "Subject Through Glass" in the Subject Matter section
		subject is.
	Inside Vehicle	A video of an enclosed compartment of a moving vehicle taken from inside that compartment. Examples are a video of a driver of a car taken by that car's passenger, videos of other passengers in an airplane, train, etc.
	Vehicle, Out	A video of an outdoor scene taken from an open moving vehicle. Examples are, videos taken from a convertible, roller coaster, boat deck, etc.
	Underwater/Pool	Taken underwater in a swimming pool
	Underwater/Scuba	Videos taken underwater in open natural bodies of water while snorkeling or scuba diving
	Water Park Rides	Videos taken where the videographer is on the water rides
TRADITIONAL SUBJECT MATTER	**Baby**	Infant, 12 months or younger
	Child	People who appear to be 18 years old or younger.
	Adult	People who appear to be older than 18 years of age.
	People: One Person	The primary subject includes only one person.
	People: Group of Two	The primary subject includes two people.
	People: Group of Three or More	The primary subject includes three or more people. This description applies to the primary subject and not to incidental people in the background.
	Crowd	The primary subject includes a large number of people in the distance.
	Animals	Pets (e.g., dogs, cats, horses, fish, birds, hamsters), wild animals, zoos and animal shows. Animals are generally 'live' animals, not dead. Those stuffed or mounted (taxidermy) may qualify depending on how "lively" they look.
	Buildings/City/Structures	General views of cities or buildings, and videos where a

	person's home is a primary subject of the video. It also includes videos of towers and other man made structures, such as signs, tunnels, roads, and amusement rides.
Nature/Landscape	Landscapes, flowers (including home flower gardens), supplants or cut flowers (if they are the Primary Subject), trees, various foliage, and sunsets.
Sunset (Sunrise)	the sun needs to be in front of the camera (though not necessarily in view)
Desert	fairly open desert view, with little or no trees/grass
Mountain	open, whole mountains (not just rocks), mid-range view
Field	fairly open natural landscape (not cluttered with trees and plants)
Forest	pictures of primarily trees
Snow Scenes	Snow is a significant part of the video. These are recorded because they could possibly affect the printer.
Beach/Other Water/Etc	Largely made up (1/3 of the frame or more) of a sandy beach or some body of water (e.g., swimming pool, lake, river).Note "beach" should be explicitly called out. In a more strict definition, a "beach" scene contains at least 10% each of water, sand, and sky, and was taken from land. Pictures taken primarily of water from a boat should be called "open water".
Urban	cityscape (tall buildings, at least one)
Suburban	pictures of houses and yards
Street	urban, with pavement in view
Highway	open view of high way, with substantial pavement surface
Documentation	These are usually videos taken for record keeping or insurance purposes (e.g., diamond ring, land property, new construction, a stamp collection, figurines, and hunting or fishing "trophies"

	such as deer, wild game or a string of fish). Do not record the actual item being videoed, only the fact that it is documentation.
Exhibits/Displays/Etc.	Objects placed in public view for deliberate showing (e.g., an auto show, museum exhibits, art, a posted menu, and other static items).
Furnishings	Furniture inside a person's home.
Toys/Hobbies	Children playing with toys, people building a model, matchbox cars, etc.
Shows/ Etc.	rock concerts, plays, and other stage events.
Parades	
Cake	Birthday or some other special occasion cake.
Presents/Gifts	Clues are the presence of wrapping paper, gifts displayed in boxes, etc.
Christmas Tree	Evergreen tree (real or artificial) with Christmas decorations (indoors or out).
Car/Van/Etc.	A vehicle which is part of the subject or is the primary subject. This also includes trains, boats (in water), airplanes, bikes, individual cars of an amusement ride, etc.
Subject Through Glass	Video taken through a window, or video of a framed painting protected by glass cover. These types of videos are identified by careful observation of mirrored images or dirt on the glass.
Other	Anything that does not fit into the above 20 categories (e.g., television, masks, decorations, computer screens, food).
Fall Foliage	Secondary Subject Matter – Call Fall Foliage whenever turned leaves on tree (e.g., excluding a close-up of one leaf) are a significant part of the scene, even if it is not part of the primary subject
Green Foliage	Secondary Subject Matter – Call Green Foliage whenever it is a significant part of the

		scene, even if it is not part of the primary subject.
	Sky	Secondary Subject Matter – Call sky whenever it is a significant part of the scene, even if it is not a part of the primary subject.
	Old Photographs (or Old Photo)	Use for old photographs
	Border	Use when image has had a border added.
	Text	Use when any text has been added to the image. This includes date stamps.
	Black and White	Use for black and white images.
	Other Digital Creation	
	Video Portrait	When the camera is held on a person's face in order to capture a 'video still' of them. For example, when the person video taping is going around the room and purposely captures each person for a couple of seconds to record their video 'portrait'.
	Variety of Skin Tones	The presence of more than one of the many flesh tones (Asian, African or Caucasian, etc.)
OCCASION	Around the house	Typically these videos are taken inside the home or in the yard. If video takers leave home, the videos become "Other Special Occasions".
	Amusement Park Visit	
	Birthday Party	
	Ceremony-Grad	This event is typically portrayed with a birthday cake, balloons, and birthday caps.
	Ceremony-Religious	
	Ceremony-Other	
	Childhood Moment	Videos of ceremonies excluding weddings. These would include baptisms, graduations (caps and gowns), Bar Mitzvahs, etc.
	Day Trip	
	Hiking	One day Mini-vacations taken to locations near by
		locations such as Niagara Falls, a theme park, Wineries, etc. There is no change of clothes, luggage or other indications that the trip lasted more than a day.
	Holiday-Christmas	
	Holidays - Other	Videos are of a Christmas tree and the usual Christmas decorations, but they are not necessarily taken on Christmas Day.
	Other Occasions	Videos of any holiday other than Christmas. Halloween, Easter, Fourth of July etc. Can include decorations for specific holidays not taken on the specific day of the holiday.
	Pet Moment	This includes school plays, professional sporting events and shopping. It also includes any other occasion that people have left home to attend.
	Party-BabyShower	
	Party-BridalShower	
	Party-Graduation	
	Party-Pool	
	Party-Other	
	Picnic	This includes parties other than a birthday party, such as a group of people with drinks and munchies.
	Playground Visit	The video is taken outdoors, with or without a picnic table, indoors (shelter/pavilion) or outdoors (probably should separate the two), people and food in view.
	Recreation	
	Renovation	This includes videos of people engaged in amateur recreational activities such as skiing, tennis, golf, and fishing. The videographer is capable of participating.
	Sport-Soccer	
	Sport-Football	
	Sport-Baseball/Softball	

	Sport-Basketball	
	Sports - Volleyball	
	Sport-Skiing	
	Sport-Skating	
	Sport-Tennis	
	Sports - Golfing	
	Sport-Swimming	
	Sport-Track	
	Sport-Field	
	Sport-Bowling	
	Sport-Other	
	Travel	
	Vacation	
	Visiting other Homes	This includes trips with an overnight stay, but does not include one day trips to nearby locations. Clues include luggage and change of clothes.
	Wedding	Videos taken in homes other than the home of the videographer. Some indication of travel to another location must be included in the roll of videos to use this category.
Audio	Applause	
	Background Talking	
	Camera Motor for Zoom	

	Cheer	
	Conversation (including chatter)	
	Formal Speech	
	Laughter	
	Music	
	Narration	
	Silence	
	Singing Audio	
	Traffic	
	Unrecognizable Sound	
	Water	
	Whistle	
	Wind	
	Other Sound	
	Fast Pan	
	Steady Pan	
	Following/Tracking Moving Subject	
Camera Motion	Camera Still (Handheld)	
	Camera Still (Tripod)	
	Camera Tilt	
	Camera Zoom	