MovieBase: A Movie Database For Event Detection And Behavioral Analysis

Tat-Seng Chua[1], Sheng Tang[1,2*], Remi Trichet[1], Hung Khoon Tan[1], Yan Song[1,2*]
[1]School of Computing, National University of Singapore, Singapore, 117417
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 100190
chuats@comp.nus.edu.sg, ts@ict.ac.cn, dcsrt@nus.edu.sg, hktan@cs.cityu.edu.hk, songyan@ict.ac.cn

ABSTRACT

The overwhelming amount of multimedia entities shared over the web has given rise to the need for semantic identification and classification of these entities. Numerous research efforts have tackled this problem by developing advanced content analysis techniques as well as leveraging on readily available tags, scripts, and blogs related to these multimedia entities. However, in many cases, especially for event detection and action recognition, the research efforts were hampered by the lack of large scale publicly available benchmarks. To address this problem, this paper presents a large-scale movie corpus named MovieBase that covers full length feature movies as well as huge volume of movie-related video clips downloaded from YouTube. The corpus is designed for research in event detection and action recognition. It offers over 71 hours of videos with a total of 69,129 shots. The corpus has been hand-labeled according to 7 audio and 11 visual concept tags to semantically define 11 event categories under the romantic and violence scenes. The corpus comes with a set of pre-extracted low-level visual, motion, audio as well as high-level features. Related results are furnished as a baseline for the movie event detection task.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Systems—multimedia databases
H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Performance, Design, Experimentation, Standardization

Keywords

Database, YouTube, Video content analysis, Video annotation, Video indexing and mining

*This work was done while they were working in National University of Singapore.

WSMC'09, October 23, 2009, Beijing, China.

1. INTRODUCTION

Recent advances in computing, storage, and video recording technologies, as well as the worldwide development of digital data exchange standards, have given rise to a huge amount of video footages on the Web. The phenomenon is vouched by the soaring activity of websites like YouTube[1] or "Flickr[2], offering users' access to millions of videos. As a matter of fact, recent estimation indicates that the YouTube website alone contains over 45 millions of user-contributed video sequences, with 65,000 new videos being added every day [1]. Moreover, these user-uploaded videos generally come with user-labeled tags that enrich the visual materials with meaningful semantic information.

The explosion of multimedia data has fueled a growing need to semantically organize and interpret this huge volume of data. Despite numerous efforts, finding semantically related videos remain challenging as little advancement has been made on this subject over the past decades. In order to bridge this well-known semantic gap, a flurry of research activities has been carried out in some related areas, such as event or region-of-interest (ROI) detection during the past few years. But the scientific community is still in need of a large-scale common benchmark for these tasks.

Over the last few years, several databases have been assembled and made available [2-4]. However, these are generally small scale databases, comprising mostly of short video clips [2-3] designed for specific tasks like traffic monitoring and surveillance [5-6]. Although relatively small, they are addressing the lack of international benchmark for videos and have contributed significantly to the development of their respective field of research. In short, large-scale real-world reference benchmarks still remain scarce. This is partly because of the amount of resources required to set up a comprehensive video dataset, which demands large storage capacity and manpower to go through the tedious and complex data selection, cleaning and annotation process.

This paper describes our effort to set up a large-scale movie video corpus named MovieBase that covers a combination of full-length feature movies as well as movie-related video clips crawled from YouTube. To set up MovieBase, we gathered videos from two different sources. The first source comprises 10 full-length offline DVD feature movies, including well-known action or romance movies like 'Resident Evil'; 'Crouching Tiger Hidden Dragon'; "Eyes Wide Shut" etc. These movies were divided into 49 video sequences, totaling 20 hours and 52 minutes in length. The second source comprises 10 films and 431 movie clips (for a total of 45

[1] http://www.youtube.com

[2] http://www.flickr.com

hours) downloaded from YouTube. These two sets of movies/videos are then combined to form a large data set. Each video sequence is divided into shots, which are commonly used as the basic unit for video analysis. Altogether, there are 69,129 shots in the corpus.

As most users are interested in movie events related to romance and violence, especially in relation to movie stars, hence we focus on tagging concepts related to these two topics in this corpus. Following studies of industries and users' needs, we label each shot according to 7 audio classes (sounds of gunshots, blunt objects, swords/knives, fists, music, explosion, and moving vehicles) and 11 visual classes (kissing, hugging, fists/arms fight, gun shooting, blunt objects, hurled objects, martial arts, knives/swords, strangling, explosion, and high speed vehicle chase). Because of limitation of resources, no spatial or temporal localization of the labeled actions in the shot is provided. We also label two sex-related classes (visual sex & sounds of making love) on additional 12 hours of 37 pornography videos which are not suitable for publication and are thus not included in MovieBase.

In order to facilitate the use of the corpus for training and testing, we provide a set of pre-extracted features. They are: (a) visual features: color histogram, edge histogram, color moments, wavelets texture and visual keywords; (b) motion features: KLT optical flow and the emerging cuboids; (c) audio features: MFCC, short-time energy, shot-time average zero-crossing rate, sub-band short-time energy, and sub-band short-time energy ratio; and (d) high-level features: face and CU-VIREO-374 concepts [7], which is a subset of LSCOM concept hierarchy. To provide a baseline of classification performance, we also provided a visual feature based baseline for event detection along with this corpus.

The dataset can be down-loaded from our Web site at: http://lms.comp.nus.edu.sg/research/MovieBase.htm. From the Web site, users can down-load all metadata derived including the shot boundaries, the ground-truth, and the extracted set of features as stated in Section 4. Because of copyright, users are expected to obtain the movies from DVD or YouTube links provided.

The potential research directions that can be pursued on this database include: (a) action recognition, behavioral recognition, or event detection; (b) video summarization; (c) visual attention modeling; and (d) concept annotation and retrieval. The contributions of this research are twofold:

- This is the largest movie/video test corpus addressing the semantic analysis issues in movies. This database presents both offline full-length featured movies and web crawled movie footages, thus providing different variety and quality (resolution) of videos.
- We provide the ground truth for 18 semantic tags at shot-level, covering most romantic and action oriented events in movies. These tags can be used for research in event detection, as well as behavioral recognition.

The remainder of this paper is organized as follows. Section 2 surveys existing multimedia datasets. Section 3 describes the characteristics of our movie corpus in terms of statistics, ground truth and possible applications. Section 4 lists and briefly discusses the low-level and high-level features extracted for use with this corpus, and Section 5 and 6 presents our training and testing sets, and some related results as baseline. Finally, Section 7 presents our conclusion and future enhancements.

2. RELATED WORK

This section reviews existing large-scale image/video datasets, and assesses the reliability of the associated ground truth or tags as well as their influence on related multimedia tasks.

2.1 Image Datasets

The completeness of the existing databases as shown in Table 1 can be evaluated according to their size, the diversity of the included concepts and the accuracy of the associated tags.

Despite their size in comparison to the more recent image sets, the well-known Corel[8] and Caltech-256[9] corpuses were specifically constructed to be diverse. For instance, each image of the Caltech-256 corpus has been sieved (according to the resolution and the semantic correspondence to the 256 concepts) by 4 different users before being accepted. Dataset diversity aspect can also be guaranteed by using concept lexicons like LSCOM (Large Scale Concepts Ontology for Multimedia)[10] and/or WordNet[11]. Along this line, ImageNet[12] was constructed upon the hierarchical structure of WordNet, achieving an impressive organization of huge amount of image data.

Ideally, tagging should be performed through hand-labeling to ensure the quality of the tags. But hand-labeling is a tedious task that requires considerable efforts that are proportional to the dataset size. Consequently, in practice, tags are more likely to be downloaded together with their images from multimedia sharing websites such as Flickr. However, the user-contributed tags in social network sites tend to contain a lot missing or spurious tags. Hence, for consistency reason, a simple manual tag completion and denoising phase have to be carried out before the data set is released. Unfortunately, due to the size of the dataset, such manual step does not guarantee the correctness and completeness of the resulting label set. In order to lighten the burden of manual labeling, an interesting alternative was proposed while designing the ESP dataset[13]. In this corpus, tags are set via an online game where players have to compete in matching the maximum possible number of words and images within a time limit. Although the labels are not detailed and often limited to common ones, an excellent quality of labeling can often be achieved.

By considering these three criterions, NUS-WIDE[14] and ImageNet[12] are the most complete image databases at the moment as shown in Table 1.

2.2 Video Datasets

Because of the tremendous size and difficulties in tagging the ground truth, video corpuses are rarer than image ones. Moreover, the temporal dimension entails a higher variability of data.

Perhaps the most renowned database for video and image content analysis and retrieval is the TRECVID[15] dataset as shown in Table 2. Beyond TRECVID, the great majority of video datasets are designed for specific applications. For instance, for the topic of tracking user's interests, a video dataset comprising 20,000 video sequences were gathered by Liu et al. [3] from the YouKu[3] website. More recently, Yang et al. [2] set up a corpus comprising 11,333 videos annotated with 11 classes for video categorization.

The increasing diversity and reliability of object tracking methods usher in the development of some dedicated datasets, like

[3] http://www.youku.com

Table 1. Main Related Works on Image Dataset

Image Dataset	Size	Characteristics	Main Use
Corel	68,040	One-label per image	Image classification
Caltech-256	30,608	256 object categories, >80 images per category	Object-oriented image retrieval
ImageNet	~3,200,000	One-label per image, hierarchically organized in 5,247 concepts based on WordNet	Object-oriented image retrieval and annotation
NUS-WIDE	269,648	Crawling from the Flickr, Multi-label with 5,018 tags, 81 manually tagged semantic concepts	Object-oriented, Concept annotation

Table 2. Main Related Works on Video Dataset

Video Dataset	Size	Characteristics	Main Use
TRECVID HLF/Search Task	~100-300 hrs /year	News videos (2003- 2006), Documentary videos (2007-2008, low quality)	Concept annotation, and video retrieval
TRECVID08 ED Task	100 hrs (5 cameras × 2 hours ×10 days)	Airport surveillance data	Surveillance event detection
TRECVID06-08 Rushes Task	~40 hrs/year	BBC dramatic rushes, unedited	Rushes video summarization
Weizmann human action dataset	90 videos, 9 subjects, 10 actions	Actions under unrealistic conditions	Primitive human action recognition
KTH human motion dataset	600 video files, 25 subjects, 6 actions, 4 scenarios		
HOHA Dataset	20.1 hours, 69 movies, 3669 clips, 12 actions, 10 scene classes	Realistic movie video clips	Most classes are for primitive human actions

VIVID[5] or VISOR[6] which caters for research in human action detection, such as walking and running. However, these datasets are not meant to provide an international benchmark for performance comparison since they merely foster research for a given application like traffic monitoring or surveillance. In short, the coverage of current object tracking databases is still limited to application-specific domains and a comprehensive dataset is still needed to include more generic and realistic settings.

In the scope of social networking, some large scale video corpuses composed of randomly extracted footage over a pre-determined period exist[4]. But the usefulness of these corpuses for computer vision research remains unclear.

Several databases have been dedicated to action recognition, like the KTH human motion dataset[16], and the Weizmann human action dataset[17]. But these datasets are not based on realistic settings as the depicted actions in these datasets are performed under perfect, unrealistic conditions; they usually show a character evolving in front of a uniform background, like a white wall.

Laptev et al [18] [28] constructed a movie dataset named HOHA by collecting clips from 32 well-known movies, making it the closest to our dataset. The training set contains two parts: one is manually annotated and the other is the results of automatic clip retrieval and annotation based on the movie scripts. According to their experiments, only 60% of the automatically labeled data was correct [18]. Each sample in this dataset is annotated according to 12 classes such as: AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, Standup. As most of the defined classes are for primitive human actions, this dataset is not suited for identifying highlights and events in movies which are of direct interests to most general users.

3. DATASET CHARACTERIZATION

This section introduces the MovieBase dataset. We will detail the motivation and procedure of building up dataset as well as statistical description of the contents.We will also briefly describe the potential research domain for which this database can be used.

3.1 Ground-Truth Setup

After investigating several movie companies and many movie watchers, we found that an overwhelming majority of movie watchers tend to be interested in movie highlights involving violence such as fighting, car chasing, gunshot, etc., and romantic events such as kissing and hugging. Thus, we define 11 event categories under the romantic and violence topics. The romance topic covers 2 categories of kissing and hugging, while the violence topic includes 9 categories on fists/arms, guns, blunt objects (clubs/bats), hurled objects, martial arts, knives/swords, strangling, explosions, and car chases.

In addition, audio information is an important complement to visual signals since sound-based and visual-based analysis of a given scene may result in different semantic classifications. For instance, slow paced scenes with visual explosion often is not accompanied by any discriminative sound, whereas scenes from a war movie may have the sound of explosions and gunfire in the background although nothing salient is occurring in the visual content. Thus in order to cover all possible angles, we include 7 audio categories (for gunshots, blunt objects, swords/knives, fists, music, explosion, and moving vehicles).

In an attempt to increase the variability of movie data, our dataset includes both offline and web downloaded videos.

The offline videos include 10 well-known full length DVD features movies. The movies are: Crouching tiger, hidden dragon; 3:10 to Yuma; Ronin; Lethal Weapon 4; Kill Bill Volume 1 & 2; Brave heart; L.A. Confidential; Resident Evil; The Beach; and Eyes Wide Shut. The movie set has a total duration of 20 hours and 52 minutes with an average resolution of 480×720 pixels. The movie set is divided into 49 video sequences according to the movies' original VOB settings.

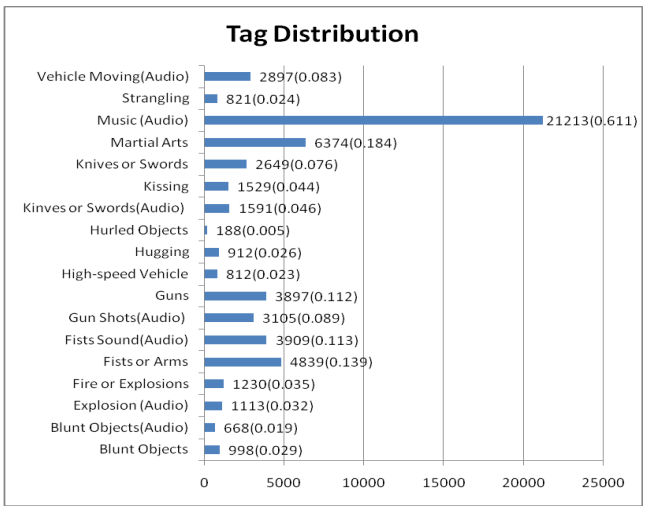

Figure 1: The distribution of tags in the dataset. The number at the end of the bar indicates the number of positive samples in each category tag. The number in bracket gives the ratio between the number of positive samples in a tag category and the total number (34,730) of positive samples in all categories.

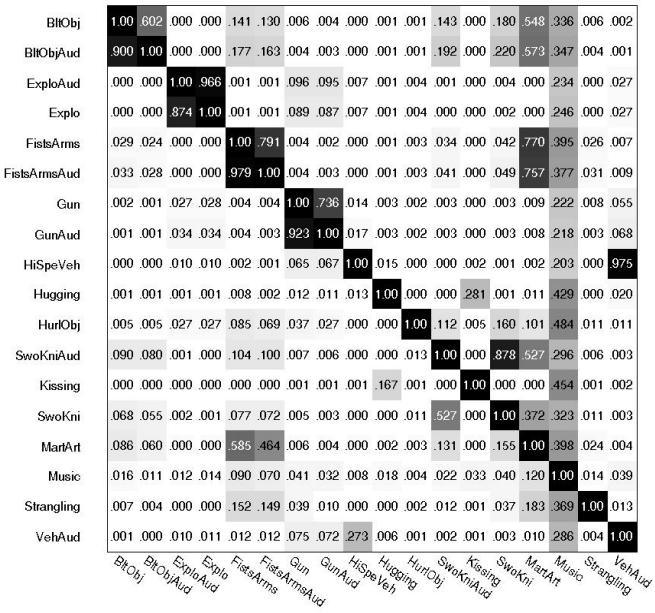

Figure 2: Pair-wise tag co-occurrence matrix

To supplement the full-length movies, we also collect relevant video clips crawled from YouTube to gather more positive video samples. These web video clips with user-defined concept tags help to lighten the manual labeling burden. Overall, the web-downloaded video set from YouTube comprises 10 films and 431 movie clips. In general, the video sequences were selected based on two criterions: (a) they are part of the movies renowned for the types of scenes that we are seeking; and (b) they are the outcome of search requests on the YouTube website by using one of our concept tags (such as: gunfight, strangling, explosion, etc.). To retrieve the list of candidate movie titles, the top 50 movies that fall under the genres of 'Action', 'Thriller', 'Sci-Fi', 'Adventure', 'Horror', 'Fantasy', 'War', 'Romance' and 'Drama', are enlisted from the IMDB website (www.imdb.com). Using this set of movie titles, we search YouTube using the following queries:

- <movie title>
- <movie title> + clip
- <movie title> + scenes
- <movie title> + <tag/its synonyms>
- <movie title> + <tag/its synonyms> + scene
- <tag/its synonyms>
- <tag/its synonyms> + scenes

For each video clip in the returned list, the potential clips are manually viewed for the presence of tags. Positive clips are then downloaded, and repeating clips are removed through manual inspection. This set consists of 45 hours of video footages with an average resolution of 240×320 pixels.

Each video is temporally segmented into shots according to a multi-resolution based shot boundary detection algorithm[19]. Shots whose durations are shorter than 1 second are removed from the dataset as they are generally difficult to be perceived by the human annotators. The resulting 69,126 shots are further analyzed to extract the corresponding key frames and the set of low-level features to be detailed in next Section.

Next we engage a group of 9 people without any a priori knowledge of the video dataset to annotate the data manually through the annotation system[20]. Each shot is labeled as positive or negative for each of the 18 tags by two different annotators. The ambiguities on the annotation are resolved by using a third annotator.

3.2 Statistical Descriptions of Database

Here we provide the statistical descriptions of the dataset, by mainly considering the differences between the tags in terms of their cardinality, repartition and co-occurrences, and the relation between the tag and its corresponding shot length.

3.2.1 Dataset Distribution

Figure 1 shows the distribution of the tags in our dataset. The "Music" tag is the most common tag covering 61.1% of all shots, whereas "Hurled Objects" is the rarest one covering only 0.5% of the shots. The numbers of the hugging and kissing positive samples amount to 912 and 1,529 respectively, much larger than those (66, 103) of the HugPerson and Kiss in the dataset[28].

The general relation between pairs of tags can be deduced from the pair-wise tag co-occurrence matrix shown in Figure 2. The element $c_{i,j}$ of the matrix is computed as follows:

$$C_{i,j} = \frac{\text{number of co-occurrences of tags i \& j}}{\text{number of occurrences of tag i}}$$

Therefore, $c_{i,j}$ denotes the conditional probability of co-occurrence of both tag i and tag j given that tag i occurs. From Figure 2, we can see that most of the visual tags tend to have high correlations with their audio counterparts as expected; and the "Fists or Arms" tends to co-occur with "Martial Arts". We also notice that "Music" is the least discriminative tag due to its relatively uniform distribution over all the other tags.

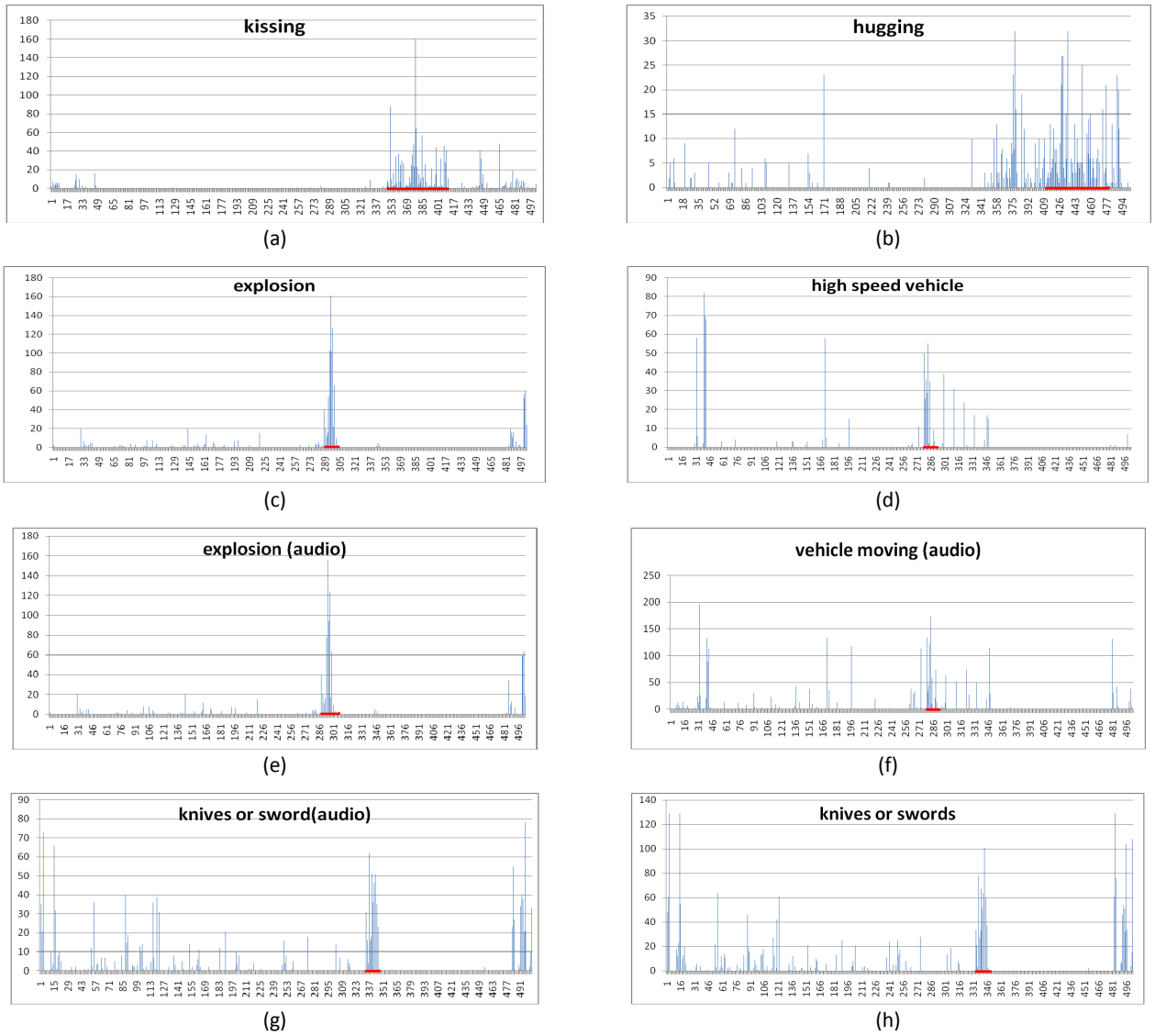

Figure 3: Frequency of individual tag over the dataset videos

The more detailed distributions of tags are shown in Figures 3(a-h) which show the frequency distribution of individual tag over the entire dataset. The horizontal axis gives the index ID of the video sequences, the source of which is illustrated in Table 3. The video sequences with IDs 1-49 and 481-503 come from full length movies; while videos with IDs 50-480 are downloaded from YouTube. The red lines along the horizontal axis in Figure 3 cover the video indexes that are downloaded by the corresponding relevant key words. It is evident from Figure 3 that the videos downloaded from YouTube by relevant key words contain adequate positive samples, and the corresponding tags also appear in common full-length movies of different genres. This demonstrates that the method we used to collect positive samples is effective.

3.2.2 Shot Length

Table 4 provides information on the distribution of shot lengths for all the 18 concept tags. It is noticed that the shot lengths of romantic tags such as "kissing" and "hugging", as well as "blunt objects", "knives" and their audio counterparts, tend to be longer than the other tags. On the other hand, "guns" and "explosions" and their associated audio tags tend to have shorter shot length as expected.

3.3 Research tasks

In general, the dataset can be applied to any video semantic analysis tasks where a spatial localization of the object/scene of interest is not required. Nevertheless, we define a set of related tasks that can be supported using this dataset.

Action recognition, Behavioral recognition or Event detection: Action or behavior recognition performs the identification of a particular human task in a given video shot. This includes primitive actions such as the person running, or composite behaviors involving a group of continuous actions covering one or more people like chasing and fighting. By defining a shot as a

Table 3. Indices of video sequences from different sources

Video Index	Video Source
1-49	10 full length movies
50-277	downloaded from YouTube by key words related to fist,gun and blunt object
278-293	downloaded from YouTube by key words related to car chase
294-307	downloaded from YouTube by key words related to explosion
308-323	downloaded from YouTube by key words related to martial arts
324-336	downloaded from YouTube by key words related to strangle
337-348	downloaded from YouTube by key words related to swords and knives
349-412	downloaded from YouTube by key words related to kissing
413-480	downloaded from YouTube by key words related to hugging
481-503	10 full length movies downloaded from YouTube and other websites where each movie has been partitioned into 2 or 3 smaller clips by uploaders

Table 4: Statistics of shot lengths according to the tag

	Max (# of frames)	Min (# of frames)	Average
Blunt Objects	5711	25	105.63
Blunt Objects(Audio)	5711	25	113.62
Explosion (Audio)	1100	26	76.7
Fire or Explosions	1100	26	78.88
Fists or Arms	5711	25	90
Fist(Audio)	5711	25	89.7
Guns	3807	8	73.73
Gun Shots(Audio)	2219	8	69.18
High Speed Vehicle	496	25	69.05
Hugging	1381	7	137.88
Hurled Object	765	25	78.3
Knives or Swords (Audio)	1697	26	92.64
Kissing	3365	6	167.59
Knives or Swords	3289	25	105.67
Martial Art	5711	12	87.07
Music	25235	12	129.97
Strangling	1549	25	85.09
Vehicle Moving(Audio)	2219	24	84.68

continuous action in space and time, this domain is closely related to event detection where the spatio-temporal patterns of the actions need to be localized in the entire video. Currently, most behavioral recognition research is performed in specific environments, such as the metro surveillance, where the view and environment are constrained in order to narrow the variability of the scenes. The absence of these constraints in movie-based dataset makes it an extremely challenging dataset for behavioral recognition.

Video summarization or Visual attention extraction: This category of applications attempts to detect the key moments of a video sequence in order to summarize the sequence. Despite the lack of spatial information, this database can be broadly used to test the reliability of such algorithms.

Video concept annotation and retrieval: The last category of application is concept annotation and retrieval especially for the event-related concepts. Although there are many databases for generalized video concept annotation and retrieval, databases that can be used for motion and event-related concept detection is still lacking. This database will be a significant addition to existing database for video concept annotation and retrieval.

4. SHOT-LEVEL FEATURES

We provide a comprehensive set of features in order to offer a common platform for fair and easy comparison, specifically, for the task of event detection. Three different categories of features are provided, based on their usefulness and potential to facilitate event detection. They are the visual-based features involving only static frames; motion-based features extracted at the shot level; and other supporting features such as the audio features, face and concept-based detectors.

4.1 Visual-based Features

The low-level visual features are mostly extracted on the keyframes (only one keyframe is extracted for each shot). The set of visual features are similar to that provided in [14] except for the visual keyword. For completeness, we provide a brief description of these features; users are referred to [14] for more details.

A) **Color Histogram**: This global representation of keyframe is based on the distribution of pixels in an uniformly partitioned LAB color space. Each of the 3 channels, L , A and B is linearly quantized into 4 bins, resulting in a 64-D feature vector.

B) **Color Auto-Correlogram**: The auto-correlogram feature characterizes the spatial correlation between pairs of identical color pixels. The HSV color components are quantized into 36 bins and the distance metric into four odd intervals, resulting in a 144-D descriptor (36×4).

C) **Color Moments**: To further incorporate spatial relationship into the color content, a keyframe is partitioned into a 5×5 grid and each patch is represented using the first three moments of the color distribution, i.e. the mean, standard deviation and the third root of the skewness of each color channel. The color moments for each patch are then concatenated to form a 255-D feature vector.

D) **Edge Histogram**: Similar to color moments, the edge histogram is computed on the 5×5 grid. At each pixel, the orientation of the gray level changes are computed and assigned to one of the 73 bins, where the edge directions are quantized into 72 bins at 5 degrees each while the 73^{th} bin is used to store the number of non-edge pixels. The edge points are detected using the Canny filter and the gradient of each edge point is determined through the Sobel operator.

E) **Wavelet Texture**: The wavelet transform is a localized texture feature where the images are decomposed using a family of basis wavelet functions. A three-level wavelet transform filters and subsamples the image recursively, and extracts the mean and standard deviation of the energy distribution of the sub-band at each level to form a 24-dimension (3×4×2) vector.

F) **Visual Keyword**: This feature is different from the features used in [14]. We use the Difference of Gaussian (DoG) detector and SIFT descriptors [21] for keypoint detection and description, respectively. A lexicon of 500 visual words is generated through k-means clustering of the 100,000 keypoints randomly sampled

from the whole data set. Given a keyframe, a soft-weighting scheme is used to weight the significance of each word in the keyframe. The soft-weighting scheme has been reported to have a superior performance as compared to the traditional TF-IDF weighting scheme [22].

4.2 Motion-based Features

The visual-based features at frame-level do not take into account the temporal cues inherent in a video medium. Temporal information has been shown to be particularly useful to filter out noisy points [23] and derive motion-based features [24] for the purpose of detecting motion-based events, such as people walking, door opening, and car moving, etc. These motion-based events would be too difficult to be reliably detected in a 2D environment without the use of motion features.

A) **Cuboids** [25]: The cuboid is a direct 3D extension of the 2D local interest points where the spatio-temporal corners are detected. The local sparse feature proceeds not only along the spatial dimensions but also in the temporal dimension. The response function has the form

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

where I is the image, g is 2D *spatial* Gaussian smoothing kernel applied *spatially* to every frame, while h_{ev} and h_{od} are the 1D Gabor filters applied to the *temporal* dimension. In general, the detector responds strongly to a range of motions, especially those involving periodic motions. For descriptors, local histograms are extracted from each cuboid as feature vectors which could then be compared through simple Euclidean distance. The cuboids are extracted at the point and shot granularity:

- At the point granularity, the flattened gradient is used as descriptors, where the cuboid is first flattened into a vector before PCA transformation is applied to reduce the dimensionality.
- At the shot granularity, a histogram of cuboid class is used. The method is similar to visual keyword generation where the cuboid class prototypes are first computed through a clustering process (*k*-means) and all cuboid descriptors are then assigned to the nearest prototype.

B) **KLT** [26]: The Kanade-Lucas-Tomasi tracker algorithm performs the tracking of some optimal features. The classical feature is used, namely a textured patch whose Eigen values of the 2x2 local intensity gradient matrix exceed a predefined threshold. The trajectories are then described by using the intra- and inter-context descriptors proposed in [27], which has reported good performance in action-recognition tasks. The intra-context descriptor models the dynamic properties of a system through an ergodic Markov chain. The inter-context descriptor further evaluates the interactions among the different features by considering more detailed information, such as the relative position of features and the local density of features by building upon the Markov stationary distribution.

4.3 Other Supporting Features

Despite the recent strides in visual recognition research, object detection remains largely an open problem. The problem is further aggravated when the event of interest is more specific but loosely defined, such as the 'fighting with a blunt objects'. In order to facilitate the detection of such events, other accompanying modalities, such as audio, face and high level feature detectors, may be useful.

A) **Audio Features:** Certain events, such as explosion, sword fighting or vehicle chase etc, exhibit distinctive audial patterns which can be tapped to enhance the performance of event detection. We divide the audio stream in shots into audio short-time frames from which 5 audios features are extracted, resulting in a 36-D audio feature vector. The features are: (a) 1-D short-time energy; (b) 1-D short-time average zero-crossing rate; (c) 4-D sub-band short-time energy; (d) 4-D sub-band short-time energy ratio; and (e) 26-D MFCC.

B) **Face**: Specialized detectors such as face recognition can be particularly helpful to infer the presence of a person and his/her corresponding actions. Given a frame, faces are detected with the following information: (a) their locations (b) scales, (c) directions; and (d) the detection scores which indicate the accuracy of face detection.

C) **CU-VIREO-**374 **concepts** [7]: We also provide the SVM predictions of the CU-VIREO-374 concept set which contains 374 semantic concepts carefully selected from the LSCOM [10] ontology. Examples of the concepts include car, explosion_fire etc. While high-level concepts have been shown to be able to narrow the semantic gaps by exploring the correlations among concepts, the usefulness of high-level concepts for event detection remains largely unexplored.

5. TESTING & TRAINING SETS

To facilitate testing, we use the first 49 offline full movie video sequences (correspond to 10 offline DVD feature movies) as the test set. The set includes 18,283 shots.

For training, we partition the remaining video sequences down-loaded from YouTube into 3 different training sets: (1) movie group covering the 10 YouTube down-loaded full-length feature movies; it includes 33,494 shots covering video sequences with ids from 481-503; (2) web-download video group with 17,352 shots, covering sequences with ids from 50-480; (3) combination of 1 & 2. The combinations enable us to train our systems using different styles of videos, namely, using full length movie as training (Training Set 1); using the YouTube down-loaded video clips as training (Training Set 2); or combination of both (Training Set 3).

6. BASELINE

As part of the corpus, we carry out the baseline runs on the test set by using the 3 combination of training sets. Based on past experience of TRECVID evaluations [20] we use SVM for classification based on the prior fusion (concatenation) of three sets of visual features, namely, color moments, edge histogram, and wavelet texture. Figure 4 shows the average precisions (AP) of all the 11 visual tags for the 3 baseline runs. The mean average precision (MAP) for all the three runs are 0.0252, 0.0173, 0.0249 respectively.

From the above results, we can see that the average precisions of the classes such as blunt objects, hurled object and hugging are very low. This, as well as the general low MAP of the 3 baseline runs, indicates that: (a) web down-loaded video clips are helpful for training; (b) merely using the keyframe-based static visual features is insufficient to represent the contents of the shots. Audio, motion and high-level features are very important for movie event detection. For example, the sounds of gunshot can be a useful clue for violence detection. They are complementary to static visual information.

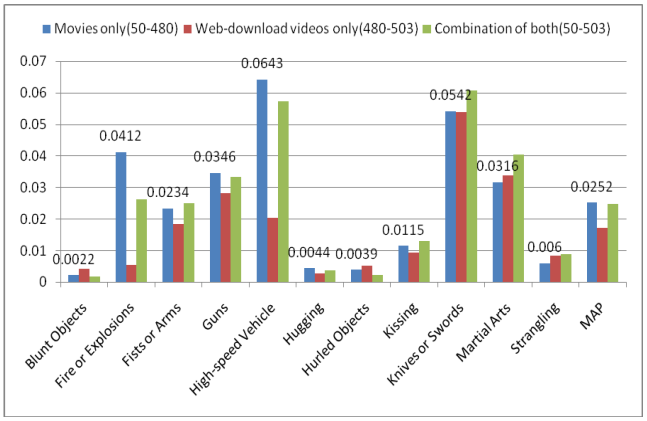

Fig.4 Average precisions on all the visual tags

7. CONCLUSION

In this paper, we introduced a large-scale movie corpus named MovieBase that covers full length movies as well as a large volume of movie-related video clips downloaded from YouTube. The corpus is designed for research in event detection and action recognition. The corpus has been hand-labeled according to 7 audio and 11 visual concept tags to semantically define 11 event categories under the romantic and violence scenes. The corpus comes with a set of pre-extracted low-level visual, motion, audio as well as high-level features. Related results are furnished as a baseline for the movie event detection task.

Several future enhancements to our corpus can be defined as follows. First, we should include text features from movie transcripts and captions available online, as well as social tagging available on YouTube. Second, we should provide further subdivision of music categories for more precise event detection. Third, for event detection and behavior recognition, we should focus on three kinds of integration in designing our learning algorithms: low-level & high-level features; static & motion-based features; and visual & audio features.

8. REFERENCES

[1] Wikipedia. http://en.wikipedia.org/wiki/youtube.

[2] L. Yang, J. Liu, X. Yang, X.S. Hua. « Multi-modality web video categorization", Multimedia Information Retrieval (MIR), pp. 265-274, 2007.

[3] L. Liu, L.F. Sun, Y. Rui, Y. Shi, S.Q. Yang. "Web Video Topic Discovery and Tracking via Bipartite Graph Reinforcement Model", In Proc. of WWW, pp. 1009-1018, 2008.

[4] X. Cheng, C. Dale, and J. Liu. "Statistics and Social Network of YouTube Videos", In Proc. Of IWQoS, 2008.

[5] R.T. Collins, X. Zhou, and S.K. Teh, "An Open Source Tracking Testbed and Evaluation Web Site", IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2005), January, 2005.

[6] R. Vezzani, R. Cucchiara, "ViSOR: Video Surveillance On-line Repository for Annotation Retrieval", in press on Proceedings of IEEE International Conference on Multimedia & Expo (IEEE ICME 2008), Hannover, 2008.

[7] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, C.-W. Ngo, "CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection", Columbia University ADVENT Technical Report #223-2008-1, Aug. 2008

[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. "Matching words and pictures". Journal of Machine Learning Research, 3:1107 – 1135, 2003.

[9] G. Griffin, A. Holub, and P. Perona. "Caltech 256 object category dataset". Technical Report UCB/CSD-04-1366, Californian institute of Technology, 2007.

[10] M. R. 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, 13(3):86–91, 2006.

[11] C. Fellbaum. WordNet: "An Electronic Lexical Database", Bradford Books, 1998.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database". To Appear in IEEE Computer Vision and Pattern Recognition (CVPR), 2009

[13] L. von Ahn and L. Dabbish. Labeling images with a computer game. In CHI04, pages 319–326, 2004.

[14] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. "NUS-WIDE: A Real-World Web Image Database from National University of Singapore", ACM International Conference on Image and Video Retrieval. Greece. Jul. 8-10, 2009.

[15] A. F. Smeaton, P. Over, and W. Kraaij. "Evaluation campaigns and TRECVID". In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321-330, 2006.

[16] Schuldt, C., Laptev, I., & Caputo, B. "Recognizing human actions: a local svm approach". In ICPR (pp. 32–36), 2004.

[17] Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. "Actions as space-time shapes". In Proceedings of the tenth IEEE international conference on computer vision (Vol. 2, pp. 1395–1402). Los Alamitos: IEEE Computer Society, 2005.

[18] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. "Learning realistic human actions from movies". CVPR, 2008.

[19] A. Chandrashekhara, H. Feng and T.-S. Chua. "Temporal multi-resolution framework for shot boundary detection and keyframe extraction". TREC (Text REtrieval Conference). Nov Gaithersburg, 2002.

[20] S. Tang, J.-T. Li, M. Li, C. Xie, Y.-Z. Liu, K. Tao, S.-X. Xu; "TRECVID 2008 High-Level Feature Extraction By MCG-ICT-CAS"; Proc. TRECVID Workshop, Gaithesburg, USA , Nov 2008

[21] D. Lowe, "Distinctive image features from scale-invariant keypoints", IJCV, pages 762-768, 2007.

[22] Y.-G. Jiang, C.-W. Ngo, J. Yang, "Towards Optimal Bag-of-Features for Object Categorization and Semantic Retrieval", CIVR 2007

[23] F. Schaffaclitzky and A. Zisserman, "Video Data Mining using Configuration of Viewpoint Invariant Regions", CVPR 2004.

[24] J. C. Nieblas, H. Wang, L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words", IJCV, 79(3), 299-318, 2008.

[25] P Dollar, V Rabaud, G Cottrell, S Belongie, "Behavior recognition via sparse spatio-temporal features", 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.

[26] J. Shi and C. Tomasi; "Good Features to Track", CVPR 1994.

[27] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua and J. Li; "Hierarchical Saptio-Temporal Context Modeling for Action Recognition". CVPR 2009.

[28] M. Marszałek, I. Laptev, C. Schmid; "Actions in Context". CVPR, 2009