Consumer Video Understanding: A Benchmark Database and An Evaluation of Human and Machine Performance

Yu-Gang Jiang[§], Guangnan Ye[§], Shih-Fu Chang[§], Daniel Ellis[§], Alexander C. Loui[†]
[§]Columbia University, New York, NY 10027, U.S.A.
[†]Kodak Research Labs, Rochester, NY 14650, U.S.A.
{yjiang, gy2179, sfchang, dpwe}@ee.columbia.edu; alexander.loui@kodak.com

ABSTRACT

Recognizing visual content in unconstrained videos has become a very important problem for many applications. Existing corpora for video analysis lack scale and/or content diversity, and thus limited the needed progress in this critical area. In this paper, we describe and release a new database called CCV, containing 9,317 web videos over 20 semantic categories, including events like "baseball" and "parade", scenes like "beach", and objects like "cat". The database was collected with extra care to ensure relevance to consumer interest and originality of video content without post-editing. Such videos typically have very little textual annotation and thus can benefit from the development of automatic content analysis techniques.

We used Amazon MTurk platform to perform manual annotation, and studied the behaviors and performance of human annotators on MTurk. We also compared the abilities in understanding consumer video content by humans and machines. For the latter, we implemented automatic classifiers using state-of-the-art multi-modal approach that achieved top performance in recent TRECVID multimedia event detection task. Results confirmed classifiers fusing audio and video features significantly outperform single-modality solutions. We also found that humans are much better at understanding categories of nonrigid objects such as "cat", while current automatic techniques are relatively close to humans in recognizing categories that have distinctive background scenes or audio patterns.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Collection*

General Terms

Standardization, Experimentation, Measurement.

Keywords

Consumer videos, database, multi-modal features, human recognition accuracy.

ICMR '11, April 17-20, Trento, Italy

1. INTRODUCTION

The explosive growth of digital videos on the Internet demands effective methods for automatic recognition and indexing of visual content. Such techniques have great potential in many important applications such as video search and open-source intelligence analysis. Key to the development of visual recognition systems is the construction of corpora with sufficient annotations for training robust models. The computer vision community has devoted several efforts to the benchmarking of image understanding tasks such as scene and object recognition. For example, LabelMe [19], ImageNet [6], and PASCAL VOC series [7] provide databases for evaluating object recognition performance, and the newly created SUN database [24] offers a comprehensive set for scene recognition research. In contrast, databases for video understanding are still quite limited in either scale or content diversity. Most video databases were designed for human action recognition, and consist of videos collected in controlled environments with clear backgrounds and little camera motion (e.g., KTH [20] and Weizmann [3]). Others are less constrained in capturing conditions, but are subject to formal or professional production processes, e.g., Hollywood movie database [13] and TRECVID databases from broadcast news to documentary [23]. A recent database from [2] focuses on human actions in videos downloaded from the Web, but does not include other content categories seen in general consumer videos.

In this paper, we introduce and release a new consumer video database called CCV[1] (Columbia Consumer Video). We are particularly interested in consumer videos because of their dominant and growing role in online video sharing. Compared to other types of online videos such as news, sports, and TV programs, consumer videos contain interesting and very diverse content, but have much less textual tag and content descriptions. For instance, on average, each consumer video in CCV has only 2.5 tags, while a general video from YouTube has 9 tags according to a recent study in [21]. The lack of textual descriptions creates significant difficulties for search engines and therefore naturally motivates research on content analysis. On the other hand, since consumer videos are captured by ordinary users without post-editing, the original audio tracks are preserved – in contrast to many news or movie videos where soundtracks are dubbed. This facilitates research on joint analysis of audio and visual features – two modalities that have been mostly studied separately in the past.

The major contributions of this paper are summarized as

[1]The database is available at www.ee.columbia.edu/dvmm/CCV.

Figure 1: Examples in Columbia Consumer Video database. The first 5 categories in the middle row are *objects* and *scenes*, and all the remaining ones are *events*. See detailed category definitions in Figure 11. Discernible faces are masked due to privacy concern.

follows:

- A large database of 9,317 consumer videos is collected from the Internet. The database is fully labeled over 20 categories, most of which are complex events, along with several objects and scenes. The categories are carefully chosen according to surveys of consumer interest and consideration of usefulness, detectability, observability, and content diversity.
- We evaluate automatic classifiers using state-of-the-art approaches that achieved top performance in recent TRECVID multimedia event detection task [11]. Carefully-designed features in both audio and visual domains, as well as their combinations, are tested.
- We study human annotation behaviors on the Amazon MTurk platform. Such studies offer interesting cues for interpreting human annotation quality. We also quantitatively measure human performance, and investigate the difference between humans and machines in consumer video understanding.

In the following we introduce the CCV database in Section 2. We then explain the Amazon MTurk annotation process in Section 3 and measure human recognition accuracy in Section 4. Section 5 includes an evaluation of automatic recognition techniques and compares machines with humans in video understanding. Finally, we conclude this paper in Section 6.

2. CCV: COLUMBIA CONSUMER VIDEO DATABASE

The first stage in constructing a video database suitable for research is to ensure the selected categories are relevant to user needs and feasible for automatic recognition. For this, we build upon our earlier effort in constructing the Kodak consumer video concept ontology [4]. The Kodak ontology contains over 100 concept definitions based on extensive user studies to evaluate the usefulness and observability (popularity) of each concept found in actual consumer videos. Some of the previous concepts, such as "wedding",

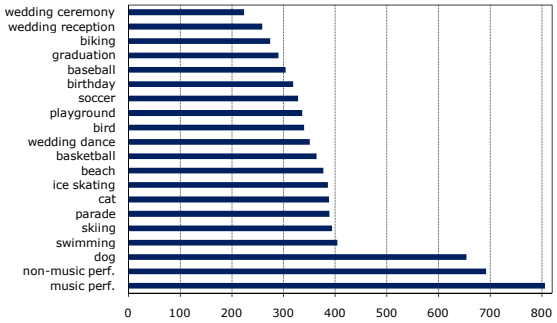


Figure 2: Number of positive samples per category in CCV, sorted in descending order from bottom to top.

turn out to be visually too diverse for computers (or even humans) to recognize. To address this, we split them into more visually and semantically distinct categories (e.g., "wedding ceremony" and "wedding dance"). Moreover, some very specific categories like "three people" and pure-audio categories such as "music" and "cheering" were dropped. This left 20 categories covering a wide range of topics including objects (e.g., "cat" and "dog"), scenes (e.g., "beach" and "playground"), sports events (e.g., "baseball" and "skiing"), and higher-level activity events (e.g., "graduation" and "music performance"). Figure 1 gives an example for each category. Detailed definitions can be found in Figure 11. It is worth noting that most of these categories are intrinsically multimodal, i.e., they can be best detected by visual and audio clues together.

Videos in CCV were downloaded from YouTube searches. We downloaded around 500 videos for each search. To ensure that only consumer videos were collected, queries were formed by combining a string "MVI" with each of the category names (e.g., "MVI and parade"). Since "MVI" is the default filename prefix for videos recorded by many digital cameras, we found it very effective in digging out consumer videos from YouTube: almost all the returns were observed to be consumer videos. The final database has 9,317 videos, with an average duration of around 80 seconds (210 hours in

total). Since the category relevance from these queries was very unreliable, we employed a subsequent stage of manual labeling to fully annotate the videos with the 20 categories. Figure 2 shows the number of positive examples per category after manual annotation, which ranges from 224 ("wedding ceremony") to 806 ("music performance"). The annotation process is described in Section 3.

2.1 Comparison with Related Databases

A few existing databases are relevant to unconstrained video analysis. We discuss the differences among these below.

Human Action Recognition Databases: Besides the simple human action video databases such as KTH [20] and Weizmann [3], there are a few unconstrained action video databases, e.g., UCF action database [14] and the Hollywood Movie database [13]. The UCF database contains 1,168 YouTube video clips of 11 human action classes such as "swinging" and "jumping". A recent extension of the UCF database contains 50 classes [2]. The Hollywood database contains 1,707 movie video clips of 12 action classes such as "kissing" and "standing up". All these databases have played a vital role in advancing action recognition research. However, with the rapid growth of user-generated videos online, there is a strong need to go beyond human actions – CCV aims at filling the gap by including a broader set of categories ranging from events, objects, to scenes, which are all culled in a rigorous manner.

Kodak Consumer Video Database [4]: The Kodak consumer videos were donated by around 100 customers of Eastman Kodak Company for research purposes. There are 1,358 video clips labeled with 25 concepts (including activities, scenes, and simple objects like "single person"). This database has been used in several recent studies, e.g., joint audio-visual modeling for video concept detection [8]. One of its critical drawbacks is that there is not enough intra-class variation. Although the videos were collected over a period of one year, the customers usually captured them under similar scene (e.g., many "picnic" videos were taken at the same location). This makes the database vulnerable to over-fitting issues. CCV does not have this limitation as it contains a much larger number of YouTube videos with more diverse content.

LabelMe Video Database [25]: The LabelMe Video database is built upon the success of the LableMe image annotation platform [19]. The authors developed an online system that allows Internet users to upload videos and label not only event categories, but also outlines and locations of moving objects. Since the labeling process is time-consuming, and does not result in any payment, it is dependent on highly-motivated users. So far, only 1321 videos have been labeled. CCV database has a larger number of videos (and can be easily expanded). Also, videos in LabelMe Video are mostly uploaded by a small number of researchers; the current edition contains many street-view videos. In contrast, CCV videos are very diverse.

TRECVID MED 2010 Database [1]: Motivated by the need to analyze complex events in large scale video collections, the annual NIST TRECVID activity initialized a new task in 2010 called Multimedia Event Detection (MED). Following the tradition of other TRECVID tasks, each year a new (or an extended) database is created for cross-site

Figure 3: Snapshot of our Amazon MTurk interface. A video is displayed on the left, and the annotator marks all the categories appearing in the video. The categories are arranged into a few groups for efficient browsing (e.g., "Celebration"). Annotation instructions are given at the top.

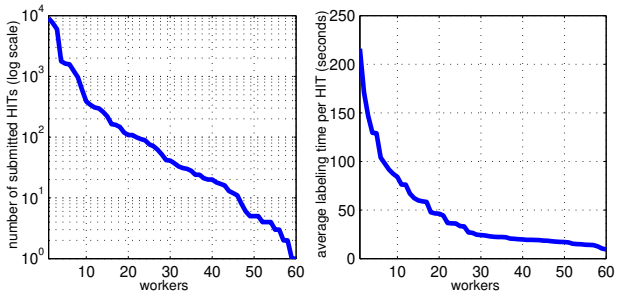
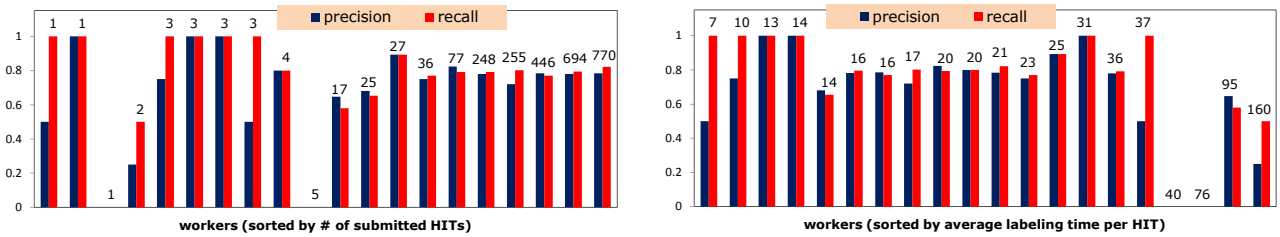


Figure 4: (Left) Number of finished HITs by each of the MTurk annotators, sorted in descending order. The Y-axis is plotted in log scale. In total there are 60 annotators engaged in this effort, with 21 of them completing more than 100 HITs. (Right) Average time per HIT spent by each annotator, sorted in descending order. 10 annotators spent more than 80 seconds (mean video duration). The others apparently did not finish viewing the entire video as instructed.

system comparisons. The 2010 edition of MED database contains 3,488 video clips (118 hours in total). Three events ("batting a run in", "making a cake", and "assembling a shelter") were used, each having around 50 positive samples in the development and test sets respectively. Compared with MED 2010, CCV is much larger and contains more positive samples per category.

3. ANNOTATION WITH AMAZON MTURK

We used the Amazon Mechanical Turk (MTurk) platform to employ Internet users as database annotators. MTurk is a crowdsourcing Internet marketplace that allows requesters to utilize human intelligence to perform customizable assignments, usually with payment. Figure 3 displays our interface on MTurk. Annotating a single video w.r.t. the 20 categories listed on the right is a single MTurk transaction, known as a HIT (Human Intelligence Task). The video playback is scheduled at 10-second segments, with the first segment automatically started and the subsequent ones triggered on demand by the user. Such arrangement is used to maximize efficiency based on the assumption that video content in the first 10 seconds is sufficient in determining some categories



(a) (b)

Figure 5: Recognition/Annotation accuracy of 20 MTurk annotators on a subset of CCV, measured in terms of both precision and recall. (a) Annotators sorted by the number of finished HITs (# HITs shown on top of each pair of bars). All the annotators who submitted more than 50 HITs achieved nearly 80% precision and similar recall rates. (b) Annotators sorted by the average time spent to finish a HIT (in seconds; shown above the bars). We did not observe strong correlation between annotation time and label accuracy.

(e.g., "baseball" and "playground"). However, annotators were encouraged to continue viewing beyond the first 10-second segment, to ensure not missing content entering the camera view in the later part. All annotators were asked to read category definitions carefully and pass a qualification round before performing the HITs. In the qualification page, the annotators had to watch 10 videos and answer several binary questions correctly. Finally, we restricted the geographic origin of the annotators to the United States to reduce linguistic confusion over category definitions.

Although the interface is carefully designed to ensure good label quality, annotators sometimes made mistakes and not all of them finished the HITs rigorously (cf. Section 4). We therefore set up two strategies to further improve label quality. First, we browsed the annotation results and inspected some suspicious labels. For example, some annotators frequently chose "I don't see any video playing". We manually checked labels from such users and rejected incorrect ones. Second and more importantly, a video was assigned to multiple independent annotators[2], then a voting strategy was used to consolidate the results and filter unreliable labels.

Figure 4 gives some statistics of the entire annotation cycle, where we show the number of submitted HITs and the mean labeling time of each annotator. Among the 60 annotators, 21 finished more than 100 HITs and a few of them labeled several thousand. On average the annotators took 22 seconds to finish one HIT. This is much shorter than the average video duration (80 seconds), indicating that they tended to finish the HITs as quickly as possible and did not watch the entire video. In the next section we analyze human annotation accuracy to determine a suitable label consolidation strategy.

4. HUMAN RECOGNITION ACCURACY

Inspired by recent work on evaluating human accuracies in image-based object/scene recognition [17, 24], we investigated the capabilities of human annotators to understand consumer videos. The purposes of this study were twofold: First, we are interested in knowing how accurately humans can recognize video content. This human recognition accuracy will be used in the next section to compare against automatic approaches. Second, this analysis makes it very easy to devise the best strategy for consolidating the labels from multiple MTurk annotators.

To measure human recognition accuracy, we randomly sampled 50 videos for each category such that each video was given that category label by at least one MTurk annotator. This forms a small subset of 896 videos (since some videos have multiple labels). We then manually labeled these videos ourselves, treating our labels as a gold-standard ground-truth. Since we understand the category definitions very well and performed the labeling tasks carefully with thorough discussions to avoid confusion, our labels are as reliable as we could reasonably hope to achieve.

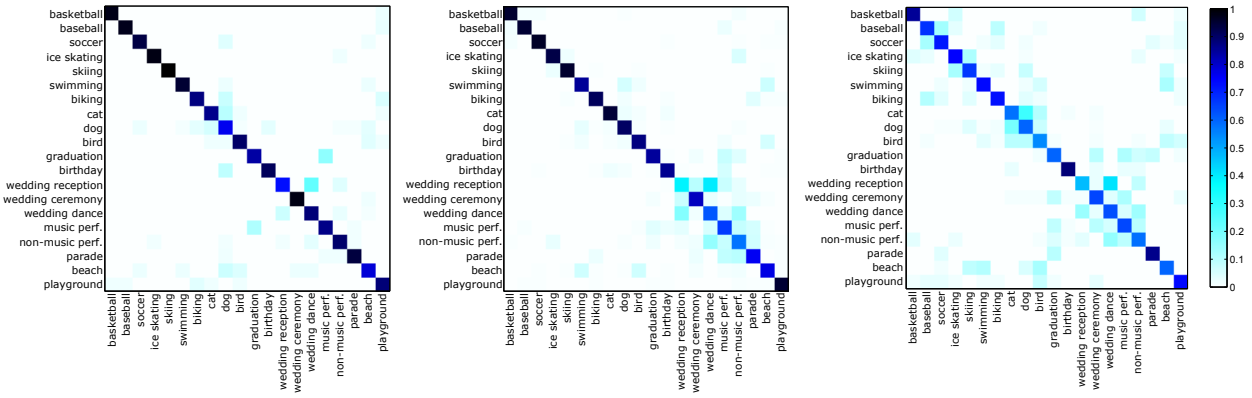
This gold-standard allows us to score the performance of the MTurk annotators. Figure 5(a) visualizes precision and recall for each of the 20 annotators with overlap on the re-labeled subset, sorted in ascending order based on the number of submitted HITs. We see that most annotators achieved very high precision and recall. Five of them are not so reliable (precision$<$ 60%), but they only finished a few HITs. The overall precision and recall are 77.4% and 79.9% respectively. Compared with the study by Xiao et al. [24] who observed that a large number of annotators had 0% accuracy, the annotators in our task are much more reliable. Apart from the task difference, this might also due to our careful process design, especially the strict qualification requirement.

To study the relationship of annotation time and accuracy, Figure 5(b) further shows the annotators' performance sorted by the time an annotator spent in finishing a HIT. Annotators who spent an average time of 10–36 seconds achieved consistent accuracy. Other annotators were not so rigorous and only finished a few HITs. From this analysis, we see no strong evidence of a correlation between annotation time and accuracy.

Figures 6(a) and 6(b) give the category confusion matrices for gold-standard labels and MTurk annotations, respectively. We observe that the 20 categories are largely exclusive, except in very few cases where two categories occasionally appear together. For instance, "wedding reception" sometimes happens together with "wedding dance", and "dog" may appear in a video of "birthday party". In contrast, MTurk annotators were sometimes confused by category pairs such as "wedding dance" vs. "non-music performance", "music performance" vs. "non-music performance", etc. This probably reflects some confusion over our category definitions.

Next we evaluate consolidation methods for merging annotations from multiple independent annotators. Here we use a voting strategy similar to [6] and evaluate the number of votes suitable to reach agreement. Figure 7 summarizes

[2]We assigned each HIT to 4 MTurk annotators.



(a) ground-truth labels (b) human recognition (c) machine recognition

Figure 6: Category confusion matrices on the CCV subset. The machine confusion matrix is computed based on classifiers using three visual/audio features (see Section 5).

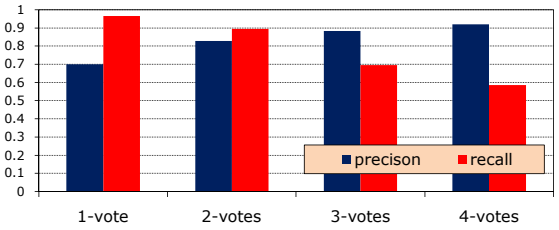


Figure 7: Accuracy of consolidated labels on the CCV subset. We use a voting strategy to combine annotations from multiple independent annotators. E.g., for method *2-votes*, a category is considered as "true" in a video if *two or more* annotators agree.

the results. As expected, as more votes are required, precision goes up while recall drops. *2-votes* emerges as the best overall compromise (precision 83%; recall 90%). Therefore we use this criterion as the consolidation strategy for the entire CCV database. To boost the label precision, we further manually filtered six categories[3] whose precisions were below 80% to remove false positives remaining in the sets returned by *2-votes*. This step took several hours but significantly boosted the label precision to about 94%.

5. MACHINE RECOGNITION ACCURACY

In this section we evaluate state-of-the-art automatic recognition methods for consumer video understanding. In particular, we focus on testing popular features and classifiers. While numerous techniques have recently been proposed for video content recognition, e.g., informative feature selection [14], contextual diffusion [10], and temporal modeling [18], we do not test them because they are all used on top of the features/classifiers for incremental performance improvement. Our aim here is to evaluate components that are critical in most visual recognition systems and assess their performance level over the new database.

We divide the database into a training set and a test set, containing 4659 and 4658 videos respectively. Positive samples of all the categories are evenly distributed into the training and test sets. The subset of 896 videos is put in the test set to facilitate comparison with human recognition accuracy. For each category, we train a one-versus-all classifier, which is then used to rank the test set according to the probability that each category appears. Performance is measured by precision-recall (PR) curves and average precision (AP; area under uninterpolated PR curve). To aggregate performance of multiple categories, we use mean AP (mAP).

[3]"Swimming", "biking", "birthday", "wedding reception", "music performance", "non-music performance".

5.1 Visual/Audio Features and Classifiers

We consider three visual and audio features that are expected to be useful and complementary for consumer video analysis.

Static SIFT Feature: SIFT has been used in almost all top-performing object and scene categorization systems. We use two sparse detectors, Difference of Gaussian [15] and Hessian Affine [16], to find local keypoints (informative image patches). Each keypoint is described by a 128 dimensional vector [15]. Since processing all frames is computationally prohibitive, we sample one frame every two seconds ($390k$ frames in total).

Spatial-Temporal Interest Points (STIP): Unlike SIFT, which describes 2D local structure in images, STIP captures a space-time volume in which pixel values have significant variations in space and time. We use Laptev's method [12] to compute locations and descriptors of STIPs. The detector was designed by extending the Harris operator to space-time [12]. Histograms of Oriented Gradients (HOG; 72 dimensions) and Histograms of Optical Flow (HOF; 72 dimensions) descriptors are computed for the 3D local volumes in the neighborhood of the detected STIPs. We use a concatenated HOG-HOF feature (144 dimensions) as the final descriptor.

MFCC Audio Feature: Besides the visual features SIFT and STIP, audio is another useful cue for understanding consumer videos. For this, we use the popular Mel-frequency cepstral coefficients (MFCC). In audio processing, the mel-frequency cepstrum (MFC) is a decorrelated representation of the short-term power spectrum of a sound. MFCCs have been widely used in many audio related applications, notably speech recognition. Although there are other recent developments on audio representation such as fingerprinting [5], MFCCs are used in this work for their simplicity and popularity. We compute an MFCC feature for every 32ms time-window with 50% (16ms) overlap.

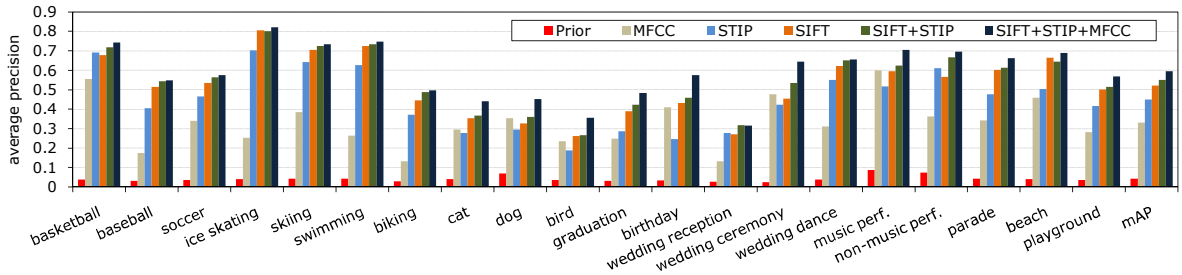


Figure 8: Machine recognition accuracy. All the three audio/visual features are effective, and their fusion produces significant gains for all the categories (best mAP 59.5%).

To convert feature sets with different cardinalities into fixed-dimensional vectors, we adopt the bag-of-X representation that has been widely used for representing documents (bag-of-words) and images (bag-of-visual-words) [22]. Given a video clip, the features extracted from each frame or audio time window are collapsed into a single bag. For SIFT, we group two visual vocabularies of 500 words for DoG and Hessian keypoints separately, and use two spatial layouts (1×1 and 2×2) to generate bag-of-SIFT histograms of 5,000 dimensions ($2 \times 500 \times (1 + 2 \times 2)$). For STIP and MFCC, we use vocabularies of 5,000 words and 4,000 words respectively. No spatial/temporal partitioning is used for either as we have found it unhelpful. For all the three features, a soft-weighting scheme is employed to alleviate the quantization effects in generating bag-of-X features [9].

With the three bag-of-X representations, video categories are learned using one-versus-all χ^2 kernel SVM, which has been proven suitable for classifying histogram-like features. Using similar features and classifiers, we achieved the best performance in the 2010 TRECVID MED task introduced earlier [11]. In the following we analyze results on CCV.

5.2 Results and Analysis

Figure 8 gives AP performance of all the three features. Comparing them individually, SIFT has the best mAP of 52.3%. It is especially good for categories with clear scene patterns such as "beach" and "ice skating". The STIP feature is not as good as SIFT, but seems more suitable for categories with moving objects under complex/diverse backgrounds, e.g., "wedding reception" and "non-music performance". MFCC audio features have the worst mAP (28.3%), but are still very discriminative, considering that the mean prior of the categories is only 4.2%. For categories with specific audio characteristics (e.g., "wedding ceremony", "dog" and "music performance"), MFCCs alone already show better performance than both visual features.

We also evaluate the usefulness of combing multiple features in Figure 8. The combination is done by average-fusion of individual SVM predictions. Other methods with adaptive fusion weights such as multiple kernel learning might further improve performance. From the results we see that fusion is always helpful. SIFT+STIP improves the mAP to 55.1% and further including MFCC audio achieves the best performance of 59.5% (8.1% relative gain over the visual features). It is interesting to notice that audio is helpful in almost every category, and for many the improvement is very significant, e.g., "dog" (↑26%), "bird" (↑33%), birthday (↑25%), "music performance" (↑13%), etc. This clearly confirms the importance of jointly modeling audio and visual features for analyzing unconstrained consumer videos, and indicates a promising direction for future research. Figure 11 (at the end of the paper) shows per-category PR curves and frames of the top 8 most confident videos in the test set.

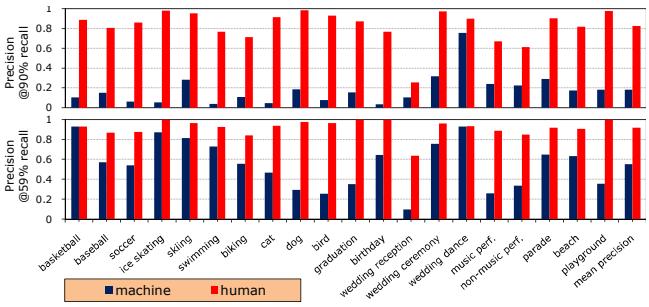


Figure 9: Human vs. machine in consumer video understanding. We compare precision at two different recall rates. Machines perform competitively for several categories at 59% recall, but are significantly below humans at a higher recall rate of 90%.

5.3 Human vs. Machine

Finally we compare humans and machines in consumer video understanding by evaluating their precision when recall rates are made equal. As seen in Figure 9, machines perform close to the human level for several categories at a moderate recall of 59%, but are far below humans at a higher recall rate of 90%. This indicates that machines can only find a portion of the true-positives with high confidence, a conclusion supported by the PR curves shown in Figure 11, where precision typically maintains a high value in the initial detection results but then plunges as more results are returned. In contrast, humans are able to recognize almost all (90%) of the relevant videos while maintaining a fairly high precision, confirming that humans are still superior at dealing with content diversity or "difficult" examples. This no doubt reflects the human ability to leverage knowledge learned over many years, far more comprehensive than the small training set used by machines. The observed performance gap between humans and machines is in line with a previous study on image annotation by Parikh et al. [17], who observed that human performance is about twice as good as machines on the PASCAL VOC database [7].

Analyzing the 20 categories separately, we see that humans are much better at recognizing nonrigid objects (e.g., "cat" and "bird") and highly complex events (e.g., "wedding reception"). Machines are relatively close to humans for categories with unique background scenes (e.g., "basketball") or characteristic audio patterns (e.g., "wedding ceremony"). Figure 10 shows some example results in the re-labeled subset. Figure 6 (b-c) further compares confusion matrices of

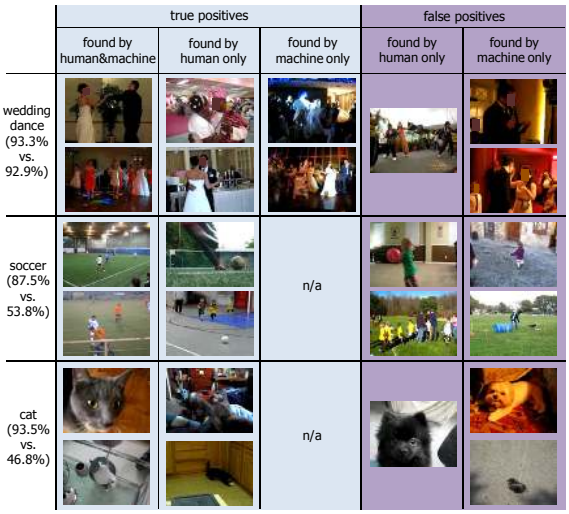


Figure 10: Example results of three categories, for which humans always perform well (top percentages) but machine performance varies (bottom percentages). Humans are able to recognize "difficult" samples (the 2^{nd} column) that computers cannot, but sometimes they also miss true positives (the 3^{rd} column) and make mistakes (the 4^{th} column).

humans and machines, indicating that machines make mistakes between categories like "cat" and "dog", "baseball" and "soccer", "ice skating" and "skiing", etc. These are probably because the current machine feature representations are not discriminative enough.

6. CONCLUSION

In this work, we constructed a new consumer video benchmark database that covers broader semantic categories than existing corpora. Based on this database, we evaluated popular machine recognition techniques and conducted a comparative study of human and machine recognition performance in video understanding. In the machine recognition experiments, we observed significant gains by fusing audio and visual features. We therefore envision a very interesting and important research topic along this direction: deep-joint audio-visual representation for video recognition, in contrast to the shallow combination used in most systems today. Through the analysis of human recognition, we found that human performance is fairly stable across all the categories. Moreover, we observed that humans are much better than machines at recognizing nonrigid objects, while current machine recognition techniques approach human's level for categories with unique background scenes or sound patterns. The database constructed by a rigorous process, the lessons learned from implementing the popular machine recognition system, and the analysis of human recognition performance will be very useful for future research on consumer video content analysis.

Acknowledgement

Many thanks to Promiti Dutta, Simon Yang, and Wei Jiang for their help/comments to this database. This work is supported in part by NGA Award # HM1582-09-1-0036, NSF Awards # CNS-07-16203 and CNS-07-51078, and a gift grant from Kodak.

7. REFERENCES

[1] TRECVID multimedia event detection track. `http://www.nist.gov/itl/iad/mig/med10.cfm/`, 2010.

[2] UCF 50 human action dataset. `http://server.cs.ucf.edu/~vision/data/UCF50.rar`, 2010.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *International Conference on Computer Vision*, 2005.

[4] A. C. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa. Kodak's consumer video benchmark data set: Concept definition and annotation. In *ACM International Workshop on Multimedia Information Retrieval*, 2007.

[5] C. Cotton and D. Ellis. Audio fingerprinting to identify multiple videos of an event. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[8] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. C. Loui. Audio-visual atoms for generic video concept classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 6:1–19, 2010.

[9] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM International Conference on Image and Video Retrieval*, 2007.

[10] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo. Domain adaptive semantic diffusion for large scale context-based video annotation. In *IEEE International Conference on Computer Vision*, 2009.

[11] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-UCF TRECVID2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[12] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64:107–123, 2005.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[14] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60:91–110, 2004.

[16] K. Mikoljczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60:63–86, 2004.

[17] D. Parikh and C. L. Zitnick. The role of features, algorithms and data in visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[18] K. Prabhakar, S. Oh, P. Wang, G. D. Abowd, and J. M. Rehg. Temporal causality for the analysis of visual events. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[19] B. Russell, A. Torralba, K. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *International Conference on Pattern Recognition*, 2004.

[21] A. S. Sharma and M. Elidrisi. Classification of multi-media content (videos on youtube) using tags and focal points. `http://www-users.cs.umn.edu/~ankur/FinalReport_PR-1.pdf`.

[22] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 2003.

[23] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *ACM International Workshop on Multimedia Information Retrieval*, 2006.

[24] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[25] J. Yuen, B. C. Russell, C. Liu, and A. Torralba. LabelMe video: building a video database with human annotations. In *International Conference on Computer Vision*, 2009.

Figure 11: Category definitions and machine recognition results including the best AP performance (in parenthesis), precision-recall curves, and frames from the top 8 detected videos in the test set (based on all the three features; ordered from left to right and top to bottom). We see significant improvements resulting from the fusion of multiple features, which achieves the best performance for all the categories. Top-ranked videos in the test set are mostly correct; false positives are marked using dashed bounding boxes. This figure is best viewed on screen with pdf magnification. Discernible faces are masked due to privacy concern.