HMDB: A Large Video Database for Human Motion Recognition

H. Kuehne
Karlsruhe Instit. of Tech.
Karlsruhe, Germany
kuehne@kit.edu

H. Jhuang E. Garrote T. Poggio
Massachusetts Institute of Technology
Cambridge, MA 02139
hueihan@mit.edu, tp@ai.mit.edu

T. Serre
Brown University
Providence, RI 02906
thomas_serre@brown.edu

Abstract

With nearly one billion online videos viewed everyday, an emerging new frontier in computer vision research is recognition and search in video. While much effort has been devoted to the collection and annotation of large scalable static image datasets containing thousands of image categories, human action datasets lag far behind. Current action recognition databases contain on the order of ten different action categories collected under fairly controlled conditions. State-of-the-art performance on these datasets is now near ceiling and thus there is a need for the design and creation of new benchmarks. To address this issue we collected the largest action video database to-date with 51 action categories, which in total contain around 7,000 manually annotated clips extracted from a variety of sources ranging from digitized movies to YouTube. We use this database to evaluate the performance of two representative computer vision systems for action recognition and explore the robustness of these methods under various conditions such as camera motion, viewpoint, video quality and occlusion.

1. Introduction

With several billion videos currently available on the internet and approximately 24 hours of video uploaded to YouTube every minute, there is an immediate need for robust algorithms that can help organize, summarize and retrieve this massive amount of data. While much effort has been devoted to the collection of realistic internet-scale static image databases [17, 23, 27, 4, 5], current action recognition datasets lag far behind. The most popular benchmark datasets, such as KTH [20], Weizmann [3] or the IXMAS dataset [25], contain around 6-11 actions each. A typical video clip in these datasets contains a single staged actor with no occlusion and very limited clutter. As they are also limited in terms of illumination and camera position variation, these databases are not quite representative of the richness and complexity of real-world action videos.

Figure 1. Sample frames from the proposed HMDB51 [1] (from top left to lower right, actions are: hand-waving, drinking, sword fighting, diving, running and kicking). Some of the key challenges are large variations in camera viewpoint and motion, the cluttered background, and changes in the position, scale, and appearances of the actors.

Recognition rates on these datasets tend to be very high. A recent survey of action recognition systems [26] reported that 12 out of the 21 tested systems perform better than 90% on the KTH dataset. For the Weizmann dataset, 14 of the 16 tested systems perform at 90% or better, 8 of the 16 better than 95%, and 3 out of 16 scored a perfect 100% recognition rate. In this context, we describe an effort to advance the field with the design of a large video database containing 51 distinct action categories, dubbed the Human Motion DataBase (HMDB51), that tries to better capture the richness and complexity of human actions (see Figure 1).

Related work. An overview of existing datasets is shown in Table 1. In this list, the Hollywood [11] and UCF50 [2] datasets are two examples of recent efforts to build more realistic action recognition datasets by considering video clips taken from real movies and YouTube. These datasets are more challenging due to large variations in camera motion, object appearance and changes in the position, scale and viewpoint of the actors, as well as cluttered background. The UCF50 dataset extends the 11 action categories from the UCF YouTube dataset for a total of 50 action categories with real-life videos taken from YouTube. Each category has been further organized by 25 groups containing video clips that share common features (*e.g*. background, camera position, *etc*.).

The UCF50, its close cousin, the UCF Sports dataset [16], and the recently introduced Olympic Sports dataset [14], contain mostly sports videos from YouTube. As a result of searching for specific titles on YouTube, these types of actions are usually unambiguous and highly distinguishable from shape cues alone (*e.g*., the raw positions of the joints or the silhouette extracted from single frames).

To demonstrate this point, we conducted a simple experiment: using Amazon Mechanical Turk, 14 joint locations were manually annotated at every frame for 5 randomly selected clips from each of the 9 action categories of the UCF Sports dataset. Using a leave-one-clip-out procedure, classifying the features derived from the joint locations at single frames results in a recognition rate above 98% (chance level 11%). This suggests that the information of static joint locations alone is sufficient for the recognition of those actions while the use of joint kinematics is not necessary. This seems unlikely to be true for more real-world scenarios. It is also incompatible with previous results of Johansson *et al*. [9], who demonstrated that joint kinematics play a critical role for the recognition of biological motion.

We conducted a similar experiment on the proposed HMDB51 where we picked 10 action categories similar to those of the UCF50 (*e.g*. climb, climb-stairs, run, walk, jump, *etc*.) and obtained manual annotations for the 14 joint locations in a set of over 1,100 random clips. The classification accuracy of features derived from the joint locations at single frames now reaches only 35% (chance level 10%) and is much lower than the 54% obtained using motion features from the entire clip (Section 4.1). We also computed the classification accuracy of the 10 action categories of the UCF50 using the motion features and obtained an accuracy of 66%.

These small experiments suggest that the proposed HMDB51 is an action dataset whose action categories mainly differ in motion rather than static poses and can thus be seen as a valid contribution for the evaluation of action recognition systems as well as for the study of relative contributions of motion *vs*. shape cues, a current topic in biological motion perception and recognition [22].

Table 1. A list of existing datasets, the number of categories, and the number of clips per category sorted by year.

Dataset	Ref	Year	Actions	Clips
KTH	[20]	2004	6	100
Weizmann	[3]	2005	9	9
IXMAS	[25]	2006	11	33
Hollywood	[11]	2008	8	30-129
UCF Sports	[16]	2009	9	14-35
Hollywood2	[13]	2009	12	61-278
UCF YouTube	[12]	2009	11	100
Olympic	[14]	2010	16	50
UCF50	[2]	2010	50	min. 100
HMDB51	[1]	2011	51	min. 101

Contributions. The proposed HMDB51 contains 51 distinct action categories, each containing at least 101 clips for a total of 6,766 video clips extracted from a wide range of sources. To the best of our knowledge, it is to-date the largest and perhaps most realistic available dataset. Each clip was validated by at least two human observers to ensure consistency. Additional meta information allows for a precise selection of testing data, as well as training and evaluation of recognition systems. The meta tags for each clip include the camera view-point, the presence or absence of camera motion, the video quality, and the number of actors involved in the action. This should permit the design of more flexible experiments to evaluate the performance of computer vision systems using selected subsets of the database.

We use the proposed HMDB51 to evaluate the performance of two representative action recognition systems. We consider the biologically-motivated action recognition system by Jhuang *et al*. [8], which is based on a model of the dorsal stream of the visual cortex and was recently shown to achieve on-par with humans for the recognition of rodent behaviors in the homecage environment [7]. We also consider the spatio-temporal bag-of-words system by Laptev and colleagues [10, 11, 24].

We compare the performance of the two systems, evaluate their robustness to various sources of image degradations and discuss the relative role of shape *vs*. motion information for action recognition. We also study the influence of various nuisances (camera motion, position, video quality, *etc*.) on the recognition performance of these systems and suggest potential avenues for future research.

2. The Human Motion DataBase (HMDB51)

2.1. Database collection

In order to collect human actions that are representative of everyday actions, we started by asking a group of students to watch videos from various internet sources and digitized movies and annotate any segment of these videos that

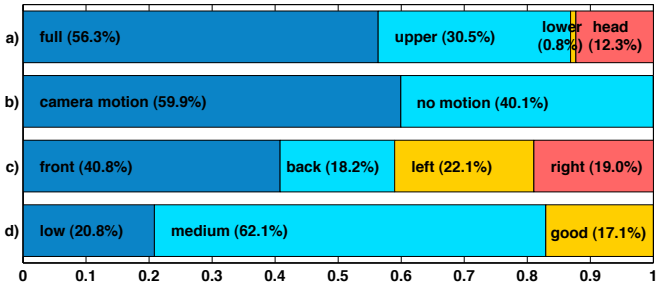

Figure 2. Distribution of the various conditions for the HMDB51: a) visible body part, b) camera motion, c) camera view point, and d) clip quality.

represents a single non-ambiguous human action. Students were asked to consider a minimum quality standard like a single action per clip, a minimum of 60 pixels in height for the main actor, minimum contrast level, minimum 1 second of clip length, and acceptable compression artifacts. The following sources were used: digitized movies, public databases such as the Prelinger archive, other videos available on the internet, and YouTube and Google videos. Thus, a first set of annotations was generated with over 60 action categories. It was reduced to 51 categories by retaining only those with at least 101 clips.

The actions categories can be grouped in five types: 1) General facial actions: *smile, laugh, chew, talk*; 2) Facial actions with object manipulation: *smoke, eat, drink*; 3) General body movements: *cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave*; 4) Body movements with object interaction: *brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw*; 5) Body movements for human interaction: *fencing, hug, kick someone, kiss, punch, shake hands, sword fight*.

2.2. Annotations

In addition to action category labels, each clip was annotated with meta information to allow for a more precise evaluation of the limitation of current computer vision systems. The meta information contains the following fields: visible body parts / occlusions indicating if the head, upper body, lower body or the full body is visible, camera motion indicating whether the camera is moving or static, camera view point relative to the actor (labeled front, back, left or right) , and the number of people involved in the action (one, two or multiple people).

The clips were also annotated according to their video quality. We consider three levels: 1) High – detailed visual elements such as the fingers and eyes of the main actor identifiable through most of the clip, limited motion blur and limited compression artifacts; 2) Medium – large body parts like the upper and lower arms and legs identifiable through most of the clip; 3) Low – large body parts not identifiable due in part to the presence of motion blur and compression artifacts. The distribution of the meta tags for the entire dataset is shown in Figure 2.

2.3. Training and testing set generation

For evaluation purposes, three distinct training and testing splits were generated from the database. The sets were built to ensure that clips from the same video were not used for both training and testing and that the relative proportions of meta tags such as camera position, video quality, motion, *etc*. were evenly distributed across the training and testing sets. For each action category in our dataset we selected sets of 70 training and 30 testing clips so that they fulfill the 70/30 balance for each meta tag with the added constraint that clips in the training and testing set could not come from the same video file.

To this end, we selected the three best results by the defined criteria from a very large number of randomly generated splits. To ensure that selected splits are not correlated with each other, we implemented a greedy approach by first picking the split with the most balanced meta tag distribution and subsequently choosing the second and third split which are least correlated with the previous splits. The correlation was measured by normalized Hamming distance. Because of the hard constraint of not using clips from the same source for training and testing, it is not always possible to find an optimal split that has perfect meta tag distribution, but we found that in practice the simple approach described above provides reasonable splits.

2.4. Video normalization

The original video sources used to extract the action clips vary in size and frame rate. To ensure consistency across the database, the height of all the frames was scaled to 240 pixels. Te width was scaled accordingly to maintain the original aspect ratio. The frame rate was converted to 30 fps for all the clips. All the clips were compressed using the *DivX 5.0* codec with the *ffmpeg* video library.

2.5. Video stabilization

A major challenge accompanying the use of video clips extracted from real-world videos is the potential presence of significant camera motion, which is the case for approximately 2/3 of the clips in our database as shown in Figure 2. As camera motion is assumed to interfere with the local motion computation and should be corrected, it follows that video stabilization is a key pre-processing step. To remove the camera motion, we used standard image stitching techniques to align frames of a clip.

Figure 3. Examples of a clip stabilized over 50 frames showing from the top to the bottom, the 1st, 30th and 50th frame of the original (left column) and stabilized clip (right column).

Table 2. The recognition accuracy of low-level color/gist cues for different action datasets.

Dataset	N	Color+ Gray+ PCA	Percent drop	Gist	Percent drop	HOG/ HOF
Hollywood	8	26.9%	16.7%	27.4%	15.2%	32.3%
UCF Sports	9	47.7%	18.6%	60.0%	-2.4%	58.6%
UCF YouTube	11	38.3%	35.0%	53.8%	8.7%	58.9%
Hollywood2	12	16.2%	68.7%	21.8%	57.8%	51.7%
UCF50	50	41.3%	13.8%	38.8%	19.0%	47.9%
HMDB51	51	8.8%	56.4%	13.4%	33.7%	20.2%

To do this, a background plane is estimated by detecting and matching salient features in two adjacent frames. Corresponding features are computed using a distance measure that includes both the absolute pixel differences and the Euler distance of the detected points. Points with a minimum distance are then matched and the RANSAC algorithm is used to estimate the geometric transformation between all neighboring frames. This is done independently for every pair of frames. Using this estimated transformation, all frames of the clip are warped and combined to achieve a stabilized clip. We visually inspected a large number of the resulting stabilized clips and found that the image stitching techniques work surprisingly well. Figure 3 shows an example. For the evaluation of the action recognition systems, the performance was reported for the original clips as well as the stabilized clips.

3. Comparison with other action datasets

To compare the proposed HMDB51 with existing real-world action datasets such as Hollywood, Hollywood2, UCF Sports, and the UCF YouTube dataset, we evaluate the discriminative power of various low-level features. For an ideal unbiased action dataset, low-level features such as color should not be predictive of the high-level action category. For low-level features we considered the mean color in the HSV color space computed for each frame over a 12×16 spatial grid as well as the combination of color and gray value and the use of PCA to reduce the feature dimension of those descriptors. Here we report the results "color + gray + PCA".

We further considered the low-level global scene information (gist) [15] computed for three frames of a clip. Gist is a coarse orientation-based representation of an image that has been shown to capture well the contextual information in a scene and shown to perform quite well on a variety of recognition tasks, see [15]. We used the source code provided by the authors.

Lastly, we compare these low-level cues with a common mid-level spatio-temporal bag-of-words cue (HOG/HOF) by computing spatial temporal interest points for all clips. A standard bag of words approach with 2,000 , 3,000 , 4,000 , and 5,000 visual words was used for classification and the best result is reported. For evaluation we used the testing and training splits that came with the datasets, otherwise a 3- or 5-fold cross validation was used for datasets without specified splits. Table 2 shows the results sorted by the number of classes (N) in each dataset. Percent drop is computed for the performance down from HOG/HOF features to each of the two types of low-level features. A small percentage drop means that the low-level features perform as well as the mid-level motion features.

Results obtained by classifying these very simple features show that the UCF Sports dataset can be classified by scene descriptors rather than by action descriptors as gist is more predictive than mid-level spatio-temporal features. We conjecture that gist features are predictive of the sports actions (*i.e.*, UCF Sports) because most sports are location-specific. For example, ball games usually occur on grass field, swimming is always in water, and most skiing happens on snow. The results also reveal that low-level features are fairly predictive as compared to mid-level features for the UCF YouTube and UCF50 dataset. This might be due to low-level biases for videos on YouTube, *e.g.*, preferred vantage points and camera positions for amateur directors. For the dataset collected from general movies or Hollywood movies, the performance of various low-level cues is on average lower than that of the mid-level spatio-temporal features. This implies that the datasets collected from YouTube tend to be biased and capture only a small range of colors and scenes across action categories compared to those collected from movies. The similar performance using low-level and mid-level features for the Hollywood dataset is likely due to the low number of source movies (12). Clips extracted from the same movie usually have similar scenes.

4. Benchmark systems

To evaluate the discriminability of our 51 action categories we focus on the class of algorithms for action recognition based on the extraction of local space-time information from videos, which have become the dominant trend in the past five years [24]. Various local space-time based approaches mainly differ in the type of detectors (*e.g.*, the implementation of the spatio-temporal filters), the feature descriptors, and the number of spatio-temporal points sampled (dense *vs*. sparse). Wang *et al*. have grouped these detectors and descriptors into six types and evaluated their performance on the KTH, UCF Sports and Hollywood2 datasets in a common experimental setup [24].

The results have shown that Laptev's combination of a histogram of oriented gradient (HOG) and histogram of oriented flow (HOF) descriptors performed best for the Hollywood2 and UCF Sports. As HMDB51 contains movies and YouTube videos, these datasets are considered the most similar in terms of video sources. Therefore, we selected the algorithm by Latptev and colleagues [11] as one of our benchmarks. To expand beyond [24], we chose for our second benchmark approaches developed by our group [21, 8]. It uses a hierarchical architecture modeled after the ventral and dorsal streams of the primate visual cortex for the task of object and action recognition, respectively.

In the following we provide a detailed comparison between these algorithms, looking in particular at the robustness of the two approaches with respect to various nuisance factors including the quality of the video and the camera motion, as well as changes in the position, scale and viewpoint of the main actors.

4.1. HOG/HOF features

The combination of HOG, which has been used for the recognition of objects and scenes, and HOF, a 3D flow-based version of HOG, has been shown to achieve state-of-the-art performance on several commonly used action datasets [11, 24]. We used the binaries provided by [11] to extract features using the Harris3D as feature detector and the HOG/HOF feature descriptors. For every clip a set of 3D Harris corners is detected and a local descriptor is computed as a concatenation of the HOG and HOF around the corner.

For classification, we implemented a bag-of-words system as described in [11]. To evaluate the best code book size, we sampled 100,000 space-time interest-point descriptors from the training set and applied the k-means clustering to obtain a set of $k = [2,000, 4,000, 6,000, 8,000]$ visual words. For every clip, each of the local point descriptors is matched to the nearest prototype returned by k-means clustering and a global feature descriptor is obtained by computing a histogram over the index of the matched codebook entries. This results in a k-dimensional feature vector where k is the number of visual words learned from k-means. These clip descriptors are then used to train and test a support vector machine (SVM) in the classification stage.

We used a SVM with an RBF kernel $K(u, v) = \exp(-\gamma * |u - v|^2))$. The parameters of the RBF kernel (the cost term C and kernel bandwidth γ) were optimized using a greedy search with a 5-fold cross-validation on the training set.

The best result for the original clips was reached for $k = 8,000$ whereas the best result for the stabilized clips was for at $k = 2000$ (see Section 5.1). To validate our re-implementation of Laptev's system, we evaluated the performance of the system on the KTH dataset and were able to reproduce the results for the HOG (81.4%) and HOF descriptors (90.7%) as reported in [24].

4.2. C2 features

Two types of C2 features have been described in the literature. One is from a model that was designed to mimic the hierarchical organization and functions of the ventral stream of the visual cortex [21]. The ventral stream is believed to be critically involved in the processing of shape information and the scale-and-position-invariant object recognition. The model starts with a pyramid of Gabor filters (S1 units at different orientations and scales), which correspond simple cells in the primary visual cortex. The next layer (C1) models the complex cells in the primary visual cortex by pooling together the activity of S1 units in a local spatial region and across scales to build some tolerance to 2D transformations (translation and size) of inputs.

The third layer (S2) responses are computed by matching the C1 inputs with a dictionary of n prototypes learned from a set of training images. As opposed to the bag-of-words approach that uses vector quantization and summarizes the indices of the matched codebook entries, we retain the similarity (ranging from 0 to 1) with each of the n prototypes. In the top layer of the feature hierarchy, a n-dimensional C2 vector is obtained for each image by pooling the maximum of S2 responses across scales and positions for each of the n prototypes. The C2 features have been shown to perform comparably to state-of-the-art algorithms applied to the problem of object recognition [21]. They have also been shown to account well for the properties of cells in the inferotemporal cortex (IT), which is the highest purely visual area in the primate brain.

Based on the work described above, Jhuang *et al*. [8] proposed a model of the dorsal stream of the visual cortex. The dorsal stream is thought to be critically involved in the processing of motion information and the perception of motion. The model starts with spatio-temporal Gabor filters that mimic the direction-sensitive simple cells in the primary visual cortex.

The dorsal stream model is a 3D (space-time) extension of the ventral stream model. The S1 units in the ventral stream model respond best to orientation in space, whereas S1 units in the dorsal stream model have non-separable spatio-temporal receptive fields and respond best to directions of motion, which could be seen as orientation in space-time. It has been suggested that motion-direction sensitive cells and shape-orientation cells perform the initial filtering for two parallel channels of feature processing, one for motion in the dorsal stream, and another for shape in the ventral stream.

Beyond the S1 layer, the dorsal steam model follows the same architecture as the ventral stream model. It contains the C1, S2, C2 layers, which perform similar operations as its ventral stream counterpart. The S2 units in the dorsal stream model are now tuned to optic-flow patterns that correspond to combinations of directions of motion whereas the ventral S2 units are tuned to shape patterns corresponding to combinations of orientations. It has been suggested that both the shape features processed in the ventral stream and the motion features processed in the dorsal stream contribute to the recognition of actions. In this work, we consider their combination by computing both types of C2 features independently and then concatenating them.

5. Evaluation

5.1. Overall recognition performance

We first evaluated the overall performance of both systems on the proposed HMDB51 averaged over three splits (see Section 2.3). Both systems exhibited comparable levels of performance slightly over 20% (chance level 2%). The confusion matrix for both systems on the original clips is shown in Figure 4. Errors seem to be randomly distributed across category labels with no apparent trends. The most surprising result is that the performance of the two systems improved only marginally after stabilization for camera motion (Table 3).

As recognition results for both systems appear relatively low compared to previously published results on other datasets [8, 11, 24], we conducted a simple experiment to find out whether this decrease in performance simply results from an increase in the number of object categories and a corresponding decrease in chance level recognition or an actual increase in the complexity of the dataset due for instance to the presence of complex background clutter and more intra-class variations. We selected 10 common actions in the HMDB51 that were similar to action categories in the UCF50 and compared the recognition performance of the HOG/HOF on video clips from the two datasets. The following is a list of matched categories: basketball / shoot_ball, biking / ride_bike, diving / dive, fencing / stab, golf swing / golf, horse riding / ride_horse, pull ups / pull-up, push-ups / push-up, rock climbing indoor / climb as well as walking with dog / walk.

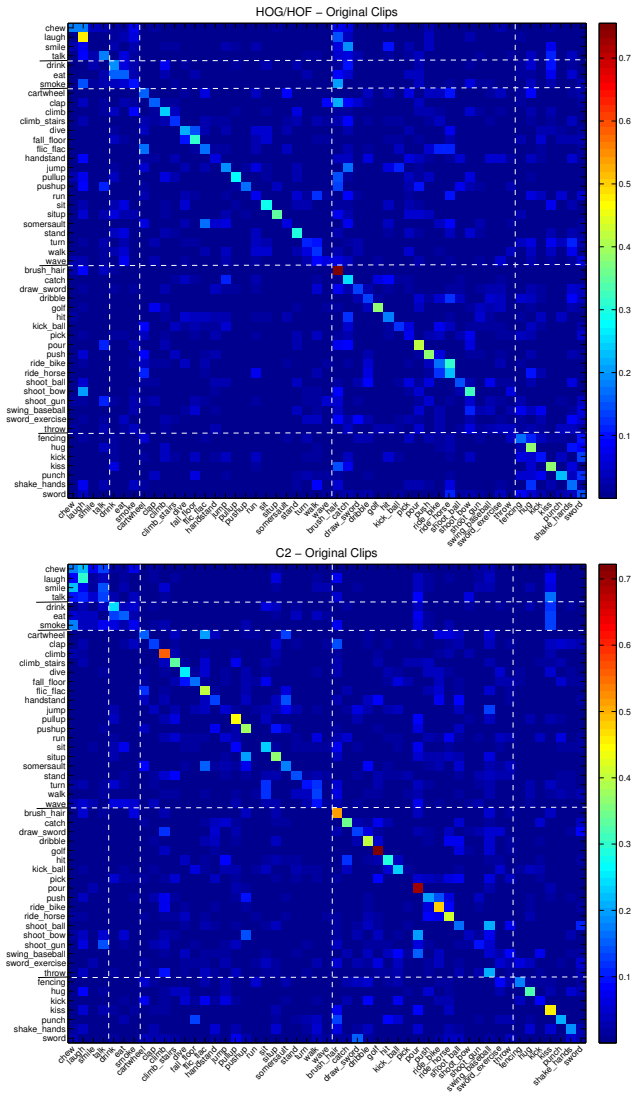

Figure 4. Confusion Matrix for HOG/HOF and the C2 features on the set of original (not stabilized) clips.

Overall, we found a mild drop in performance from the UCF50 with 66.3% accuracy down to 54.3% for similar categories on the HMDB51 (chance level 10% for both sets). These results are also comparable to the performance of the same HOG/HOF system on similar sized datasets of different actions with 51.7% over 12 categories of the Hollywood2 dataset and 58.9% over 11 categories of the UCF YouTube dataset as shown in Table 2. These results suggest that the relatively low performance of the benchmarks on the proposed HMDB51 is most likely the consequence of the increase in the number of action categories compared to older datasets.

Table 3. Performance of the benchmark systems on the HMDB51.

System	Original clips	Stabilized clips
HOG/HOF	20.44%	21.96%
C2	**22.83%**	**23.18%**

Table 4. Mean recognition performance as a function of camera motion and clip quality.

	Camera motion		Quality		
	yes	no	low	med	high
HOG/HOF	19.84%	19.99%	17.18%	18.68%	27.90%
C2	25.20%	19.13%	17.54%	23.10%	28.62%

5.2. Robustness of the benchmarks

In order to assess the relative strengths and weaknesses of the two benchmark systems on the HMDB51 in the context of various nuisance factors, we broke down their performance in terms of 1) visible body parts or equivalently the presence/absence of occlusions, 2) the presence/absence of camera motion, 3) viewpoint/ camera position, and 4) the quality of the video clips. We found that the presence/absence of occlusions and the camera position did not seem to influence performance. A major factor for the performance of the two systems was the clip quality. As shown on Table 4, from high to low quality videos, the two systems registered a drop in performance of about 10% (from 27.90%/28.62% for the HOG+HOF/C2 features for the high quality clips down to 17.18%/17.54% for the low quality clips).

A factor that affected the two systems differently was camera motion: Whereas the HOG/HOF performance was stable with the presence or absence of camera motion, surprisingly, the performance of the C2 features actually improved with the presence of camera motion. We suspect that camera motion might actually increase the response of the low-level S1 motion detectors. An alternative explanation is that the camera motion by itself might be correlated with the action category. To evaluate whether camera motion alone can be predictive of the action category, we tried to classify the mean parameters of the estimated frame-by-frame motion returned by the video stabilization algorithm. The result of 5.29% recognition shows that at least camera motion alone does not provide significant information in this case.

To further investigate how various nuisance factors may affect the recognition performance of the two systems, we conducted a logistic regression analysis to predict whether each of the two systems will be correct *vs*. incorrect for specific conditions. The logistic regression model was built as follows: the correctness of the predicted label was used as binary dependent variable, the camera viewpoints were split into one group for front and back views (because of similar appearances; front, back =0) and another group for side views (left, right =1). The occlusion condition was split into full body view (=0) and occluded views (head, upper or lower body only =1). The video quality label was converted into binary variables whereas the labels 10, 01 and 00 corresponded to a high, medium, and low quality video respectively.

Table 5. Results of the logistic regression analysis on the key factors influencing the performance of the two systems.

HOG/HOF			
Coefficient	Coef. est. β	p	odds ratio
Intercept	-1.60	0.000	0.20
Occluders	0.07	0.427	1.06
Camera motion	-0.12	0.132	0.88
View point	0.09	0.267	1.09
Med. quality	0.11	0.254	1.12
High quality	0.65	0.000	1.91
C2			
Coefficient	Coef. est. β	p	odds ratio
Intercept	-1.52	0.000	0.22
Occluders	-0.22	0.007	0.81
Camera motion	-0.43	0.000	0.65
View point	0.19	0.009	1.21
Med. quality	0.47	0.000	1.60
High quality	0.97	0.000	2.65

The estimated β coefficients for the two systems are shown in Table 5. The largest factor influencing performance for both systems remained the quality of the video clips. On average the systems were predicted to be nearly twice as likely to be correct on high *vs*. medium quality videos. This is the strongest influence factor by far. However the regression analysis also confirmed the assumption that camera motion improves classification performance. Consistent with the previous analysis based on error rates, this trend is only significant for the C2 features. The additional factors, occlusion and camera viewpoint, did not have a significant influence on the results of the HOG/HOF or C2 approach.

5.3. Shape vs. motion information

The role of shape *vs*. motion cues for the recognition of biological motion has been the subject of an intense debate. Computer vision could provide critical insight to this question as various approaches have been proposed that rely not just on motion cues like the two systems we have tested but also on single-frame shape-based cues, such as posture [18] and shape [19], and contextual information [13, 28].

We here study the relative contributions of shape *vs*. motion cues for the recognition of actions on the HMDB51. We compared the HOG/HOF descriptor with the recognition of a shape-only HOG descriptor and a motion-only HOF descriptor. We also compared the performance of the previously mentioned motion-based C2 to those of shape-based C2. Table 6 shows the performance of the various descriptors.

In general we find that shape cues alone perform much worse than motion cues alone, and their combination tends to improve recognition performance very moderately. This combination seems to affect the recognition of the original clips rather than the recognition of the stabilized clips. An

Table 6. Average performance for shape *vs.* motion cues.

HOG/HOF	HOGHOF	HOG	HOF
Original	**20.44**%	15.01%	17.95%
Stabilized	21.96%	15.47%	**22.48**%
C2	Motion+Shape	Shape	Motion
Original	**22.83**%	13.40%	21.96%
Stabilized	**23.18**%	13.44%	22.73%

earlier study [19] suggested that "local shape and flow for a single frame is enough to recognize actions". Our results suggest that the statement might be true for simple actions as is the case for the KTH dataset but motion cues do seem to be more powerful than shape cues for the recognition of complex actions like the ones in the HMDB51.

6. Conclusion

We described an effort to advance the field of action recognition with the design of what is, to our knowledge, currently the largest action dataset. With 51 action categories and just under 7,000 video clips, the proposed HMDB51 is still far from capturing the richness and the full complexity of video clips commonly found in the movies or online videos. However given the level of performance of representative state-of-the-art computer vision algorithms with accuracy about 23%, this dataset is arguably a good place to start (performance on the CalTech-101 database for object recognition started around 16% [6]). Furthermore our exhaustive evaluation of two state-of-the-art systems suggest that performance is not significantly affected over a range of factors such as camera position and motion as well as occlusions. This suggests that current methods are fairly robust with respect to these low-level video degradations but remain limited in their representative power in order to capture the complexity of human actions.

Acknowledgements

This paper describes research done in part at the Center for Biological & Computational Learning, affiliated with MIBR, BCS, CSAIL at MIT. This research was sponsored by grants from DARPA (IPTO and DSO), NSF (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05- C-7262). Additional support was provided by: Adobe, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and by the Eugene McDermott Foundation. This work is also done and supported by Brown University, Center for Computation and Visualization, and the Robert J. and Nancy D. Carney Fund for Scientific Innovation, by DARPA (DARPA-BAA-09-31), and ONR (ONR-BAA-11-001). H.K. was supported by a grant from the Ministry of Science, Research and the Arts of Baden Württemberg, Germany.

References

[1] http://serre-lab.clps.brown.edu/resources/HMDB/. 1, 2

[2] http://server.cs.ucf.edu/~vision/data.html. 2

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *ICCV*, 2005. 1, 2

[4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009. 1

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) results. http://www.pascal-network.org/challenges/voc/voc2010/workshop/index.html. 1

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *CVPR Workshop on Generative-Model Based Vision*, 2004. 8

[7] H. Jhuang, E. Garrote, J. Mutch, X. Yu, V. Khilnani, T. Poggio, A. D. Steele, and T. Serre. Automated home-cage behavioral phenotyping of mice. *Nature Communications*, 1(5):1–9, 2010. 2

[8] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *ICCV*, 2007. 2, 5, 6

[9] G. Johansson, S. Bergström, and W. Epstein. *Perceiving events and objects*. Lawrence Erlbaum Associates, 1994. 2

[10] I. Laptev. On space-time interest points. *Int. J. of Comput. Vision*, 64(2-3):107–123, 2005. 2

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, 2008. 2, 5, 6

[12] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". *CVPR*, 2009. 2

[13] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. *CVPR*, 2009. 2, 7

[14] J. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. *ECCV*, 2010. 2

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42:145–175, 2001. 4

[16] M. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. *CVPR*, 2008. 2

[17] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1):157–173, 2008. 1

[18] J. M. S. Maji, L. Bourdev. Action recognition from a distributed representation of pose and appearance. *CVPR*, 2011. 7

[19] K. Schindler and L. V. Gool. Action snippets: How many frames does human action recognition require. *CVPR*, 2008. 7, 8

[20] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. *ICPR*, 2004. 1, 2

[21] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–26, 2007. 5

[22] M. Thirkettle, C. Benton, and N. Scott-Samuel. Contributions of form, motion and task to biological motion perception. *Journal of Vision*, 9(3):1–11, 2009. 2

[23] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(30):1958–1970, 2008. 1

[24] H. Wang, M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *BMVC*, 2009. 2, 5, 6

[25] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. *ICCV*, 2007. 1, 2

[26] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Und.*, 115(2):224–241, 2010. 1

[27] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. *CVPR*, 2010. 1

[28] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. *CVPR*, 2010. 7