Heterogeneous Semantic Level Features Fusion for Action Recognition

Junjie Cai[†,*], Michele Merler[‡], Sharath Pankanti[‡] and Qi Tian[†]
[†]Department of Computer Science, University of Texas at San Antonio
[‡]IBM Thomas J. Watson Research
caijunjieustc@gmail.com,{mimerler,sharat}@us.ibm.com,qitian@cs.utsa.edu

ABSTRACT

Action recognition is an important problem in computer vision and has received substantial attention in recent years. However, it remains very challenging due to the complex interaction of static and dynamic information, as well as the high computational cost of processing video data. This paper aims to apply the success of static image semantic recognition to the video domain, by leveraging both static and motion based descriptors in different stages of the semantic ladder. We examine the effects of three types of features: low-level dynamic descriptors, intermediate-level static deep architecture outputs, and static high-level semantics. In order to combine such heterogeneous sources of information, we employ a scalable method to fuse these features. Through extensive experimental evaluations, we demonstrate that the proposed framework significantly improves action classification performance. We have obtained an accuracy of 89.59% and 62.88% on the well-known UCF-101 and HMDB-51 benchmarks, respectively, which compare favorably with the state-of-the-art.

1. INTRODUCTION

Human action recognition has been one of the challenging problems explored by the computer vision community. An action comprises a combination of semantic visual elements (people, objects, scenes) and motions, which are related to individual element movements or their interactions with each other. Therefore both the static visual information and the dynamics of the scene constitute necessary and complementary information to recognize an action. A myriad of works have been proposed to improve action recognition and classification performance [11, 21, 22], with efforts driven by the release of increasingly larger and more challenging datasets [12, 13, 36].

Out of all the proposed approaches, dense trajectories based methods with GMM codebook generation and Fisher Vector encoding have achieved competitive performances and have shown their effectiveness on this problem [11]. The motion trajectory is capable of describing subtle movements and is suitable for representing both the dynamics and appearance of a scene on a *local* level.

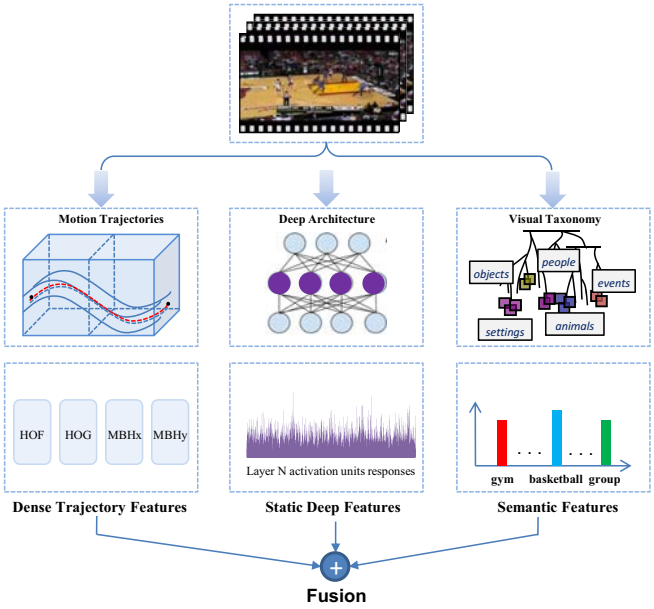

Figure 1: Overview of the proposed action recognition framework combining heterogeneous semantic level features. The semantic level of the extracted decriptors ranges from low to high as we move from left (low-level descriptors) to right (high-level semantics).

However, using only low level descriptors can not always be sufficient for action recognition purposes. This is due mostly to the gap with the semantics of an action at a global scale. Consider the example in Figure 2 showing two clips from the UCF101 dataset: for some reason the HOF descriptor of the two video clips is very similar, but on a semantic level it is quite clear that the instrument being played is completely different. Such limitation can be alleviated by the use of a set of static **Semantic** classifiers, especially in the case where some relevant concepts are part of the pool (in the example, *Grand Piano* and *Flute*).

[*]This work was done when Junjie Cai was an intern at IBM T.J Watson Research Center.

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749320.

Besides providing complementary information to low level dynamic descriptors, using sets of Semantic models constitutes also a way to transfer the knowledge accumulated from the large scale annotated training data on top of which such models were learned.

In this work, we propose to use descriptors at Higher Semantic levels in combination with the low-level dynamic ones. Figure 1 gives a conceptual view of the feature extraction process. First, we extract visual features at different semantic levels: *low-level* dynamic trajectory features (e.g., histogram of oriented gradients (HOG) [15], histogram of optical flow (HOF) [15] and motion boundary histogram (MBH) [11]), *intermediate-level* static deep features extracted from one of the intermediate layers of a Deep convolutional network trained on ImageNet, and *high-level* semantic features in the form of concatenated predictions of two sets of static visual classifiers trained from web images. The first set contains 1,000 classes from the same Deep convolutional network as the intermediate level features, the other has 1,414 classes trained as ensemble SVMs from web images.

In order to mine and fully exploit the complementary information among heterogeneous features, we explore the features relationship via a selection of fusion strategies, of which SVM based fusion proved to be the best performing.

While the use of deep features for action recognition in videos has been introduced before [3, 8], the semantic analysis to the output of the final layer of the neural network has not been yet fully explored.

The main contributions of this paper can be summarized as follows:

- We propose to use high-level static semantic classifiers to perform action recognition in videos.
- We propose a framework that jointly combines dynamic trajectory features, static deep features and high level semantic predictors for improving action classification performance.
- Extensive empirical evaluations are provided to corroborate the effectiveness of the proposed framework in detail, which achieved an accuracy of 89.59% and 62.88% on the well-known UCF-101 and HMDB51 benchmarks, respectively, which compare favorably to the state-of-the-art.

The remainder of this paper is organized as follows. After an overview of related work in Section 2, we describe the proposed action recognition framework in Section 3. In Section 4, the experimental results are presented and discussed. Finally, we conclude in Section 5.

2. RELATED WORK

The complete literature on action recognition is quite extensive and beyond the scope of this paper. In this Section, we focus on works closely related to our approach, mainly covering systems which employ features of low, intermediate, and high semantic level.

Low-level Features unlike static images, video data exhibits different views of visual patterns, such as appearance changes and motions with different boundaries, all of which play an important role in action recognition. Therefore multiple descriptors are usually extracted directly from video and each descriptor corresponds to one specific aspect of visual data. Researchers in action recognition widely made use of low-level features with BoW model. Typical low-level features employed to recognize actions in videos include histogram of oriented gradients (HOG) [15], histogram of optical flow (HOF) [15] and motion boundary histogram (MBH) [11], which are computed in local cuboids obtained around detected spatial-temporal interest points or with dense sampling schemes [16]. A combination of several features is shown to further boost recognition accuracy by leveraging fisher vector encoding. For instance, participants in THUMOS challenge [36] are encouraged to employ various kinds of dynamic features to develop novel approaches for action recognition to operate in realistic conditions [11, 34].

Moreover, a couple of works explored the integration of multiple low level descriptors to generate semantics for action recognition. Wang et al. [6] correlate interest points together and construct an action unit set to represent all actions classes with semantics. Liu et al. [22] use diffusion maps to automatically learn a semantic visual vocabulary from low-level features. In the above model, low-level trajectory features are extracted and clustered into the codebook via a generative process (e.g. GMMs), and then each feature is quantized and encoded into a high dimensional vector (e.g. Fisher Vector). However, such representation suffers from two major limitations, which are interconnected. On one hand, similarly to all other low-level descriptors, the semantic gap between the Dense Trajectories features and the nameable semantics of an action at a global scale results in a lack of interpretability of the results. This also limits the generalization power of such descriptor to other domains, which is lost in its intrinsic need to learn a codebook and Fisher vector encoding parameters for each dataset in order to obtain competitive performance. Hence, the vocabularies or units above are not discriminative and representative enough in larger video datasets, which limits their applicability in different domains.

Intermediate-Level Features Besides low-level features, recent efforts for action and event recognition in video have been devoted to mining intermediate-level semantic repre-

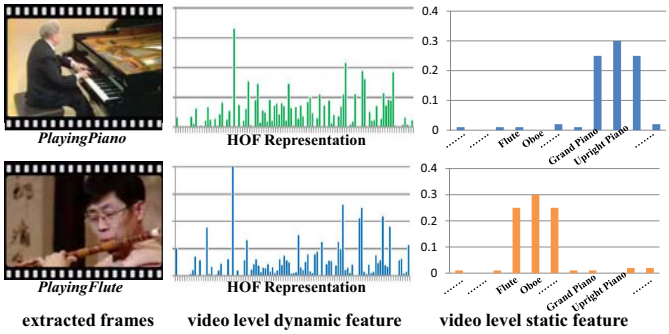

Figure 2: Comparison between a "PlayingPiano" clip and "PlayingFlute" video based on Fisher Vector HOF representation and static semantic features. In this example, the two low-level trajectory features are considered as a good match both by FV equality and Euclidean distance consistency. However, they differ significatly in static semantic space. Our proposed approach can distinguish between the two since it can overcome the shortcomings of one descriptor.

sentations. A number of researchers have been building a variety of semantic concept detectors, such as those related to people (face, anchor), acoustics (speech, music), genre (weather, financial, sports), scene, etc. [6, 20, 33, 37]. Liu et al. [23] proposed to leverage attribute-based features for action recognition. Yao et al. [24] jointly modeled the attributes (i.e. actions) and parts (i.e., objects or poselets related actions) by learning a set of sparse bases that are shown to carry much meaning. Chen et al. [31] proposed a concept discovery approach by investigating the event textual descriptions. However, most of such approaches require copious labeled data to train specifically targeted intermediate concept or attribute classifiers, which involves a costly annotation process.

Deep Neural Network There have also been some attempts to leverage a deep neural architecture for visual recognition. Deep features are generally extracted from intermediate layers of convolutional neural networks. They have been shown to set the state-of-the-art in many applications such as OCR [26], speech recognition [25] and object detection [27]. Moreover, for action recognition, Karpathy et al. [3] used raw video data as inputs instead of the hand-crafted features and compared several ConvNets architectures. Interestingly, their results indicated that the spatial temporal features learnt from 1.1M YouTube videos could not capture motion characteristics well enough to successfully generalize to other datasets. The performance turned out to be less competitive than hand-crafted trajectory-based representations. Simonyan et al. [8] further explored how to capture the complementary information from static extracted frames and motion (optical flow) via deep ConvNets.

From a semantic standpoint, the output of the hidden layer in a pre-trained neural network can be utilized to convey discriminative semantic information with respect to the data from which it was trained. One popular example in this space is the task of large scale image classification ImageNet task [9].

Our work departs from the ones reported above in two aspects. First, the dynamic trajectory features are widely used in action recognition and remain the central components of video analytical systems that generated state-of-the-art results. We aim to efficiently combine the advantages of off-the-shelf low-level dynamic trajectory features to improve the recognition performance, without discarding them completely for an independently trained deep architecture. Second, we investigate the use of semantics generated by static deep and shallow visual classifiers which provide a complementary cue to further enhance the discriminative power of action classifiers.

3. PROPOSED APPROACH

In this Section we provide a formal description of the proposed framework, starting from the details of the heterogeneous semantic level features employed and then discussing the different fusion strategies adopted to combine them.

3.1 Feature Extraction

Dynamic Dense Trajectory Features: Dynamic information is an important cue for human action understanding from video. Therefore we extract the state of the art improved dense trajectories features from input videos. To describe motion, different types of motion feature descriptors are computed in a spatial-temporal volume (i.e., spatial size of 2×2 with temporal length of 15) around the $3D$ neighborhood of the tracked points along the trajectory. Following [11], each trajectory is described by a concatenation of HOG, HOF, and MBH(along x and y directions) descriptors, forming a 396-dimensional vector (96+108+96+96).

Fisher Vector (FV) coding, derived from Fisher Kernel, was originally proposed for large scale image classification. Compared with other coding methods such as vector quantization and sparse coding, FV coding can easily obtain high-dimensional feature codes starting from a small codebook size, which has been shown to provide considerable performance improvements when utilizing linear classifiers.

Following best practices reported in the literature [7], we process the descriptors independently. For each of them, we first reduce the dimensionality by performing PCA with a ratio of 0.5. We then randomly sample 0.3 million features to learn a codebook of Gaussian Mixture Models (GMM) for each descriptor. We apply the Fisher Vector high-dimensional encoding scheme to each descriptor and the resulting super vectors are normalized by intra power normalization [35]. The normalization strategy is carried out in a block-by-block manner and each block represents the vector related to one codeword. We use $\mathbf{p^k}$ to denote a vector related to k-th Gaussian and $\|.\|$ stands for ℓ_2-norm. The normalization could be represented as $\mathbf{p}^k / \left\|\mathbf{p}^k\right\|$, where $k \in [1, K]$. Finally the normalized super vectors are concatenated to represent the motion information for a given video clip.

Static Deep Features: We leverage an existing deep learning framework as a feature extractor for video frames. Each video clip is uniformly sampled at a rate of two frames per second, and deep learning features are extracted from each frame. In order to produce a whole video clip level feature representation, we adopt a simple max pooling scheme on the individual frame descriptors over the duration of each video. We utilize the open source CAFFE [2] implementation, which is based on the deep convolutional neural network architecture by Krizhevsky et al. [4]. Since the frames in the action videos we investigate are independent from the ImageNet dataset on which the CAFFE architecture was trained, we are basically using the ImageNet model trained for previous ILSVRC image classification tasks [9] as an analog to using the prior knowledge a human obtained from previous visual experiences to learn new tasks more efficiently (in our case, action recognition).

The activations of the neurons in the intermediate hidden layers could be used as strong features for a variety of video recognition tasks because they contain much richer and more complex representations than any earlier convolutional layer in the network. In this work, we leverage as deep feature the output of the intermediate layer named with fc_6 in the CAFFE implementation. We set the network input to the raw RGB values of the frames, resized to 256*256 pixels, and the values are forward propagated through 5 convolutional layers (i.e., pooling and ReLU non-linearities) and 3 fully-connected layers (i.e., to determine its final neuron activities). We obtain the 4096-dimension vector from the intermediate fc_6 hidden layer.

The deep feature produced by this architecture has a large variation in its value distribution (i.e., [-72.8,24.8]), which is potentially problematic due to sensibility to outliers in one of its dimensions. This problem could be severe if we

consider the fact that the negative values are produced by suppressed neurons, and convey less useful cues compared with the positive ones. To address this problem, and thus produce more uniformly distributed values, we normalize each dimension using the following function:

$$f(x) = sign(x)|x|^{\alpha} \quad (1)$$

where $sign(\cdot)$ denotes the signum function and $\alpha \in [0, 1]$ is the exponent parameter. Here we empirically set α as 0.5. Finally, the feature vector is ℓ_2-normalized.

In the field of action recognition, the effectiveness of deep features has not been yet extensively studied, especially in a complementary way with dynamic trajectory features and higher level semantics. In this work, we make initial attempts on this issue, and provide feasible ways of integrating static deep features into the classification pipeline.

High Level Semantic Concepts Features: In order to richly represent the visual semantic concept space, we employ two diverse sets of static semantic classifiers which were trained on different datasets using different learning techniques.

- CAFFE1K : the set of 1,000 classifiers originating from the output of the last layer of the Deep convolutional network described in the previous Section. Each of those outputs carries one specific semantic information associated with a visual concept from the ImageNet taxonomy.

- ConceptsWeb : a set of 1,418 models trained from a set of half a million images downloaded from the web and manually annotated and organized in a hierarchical faceted taxonomy [1]. This taxonomy includes concepts related to "objects","scenes", "people", "activities" and "events". Each model SC_i is an ensemble of SVMs with linearly approximated χ^2 kernel, learned on top of bags of examples randomly sampled from the set of thousands of manually labeled web images. Each individual SVM uses one of 13 different global visual descriptors including color histogram, color correlogram, color moment, wavelet texture, edge histogram, etc., extracted at multiple different regions of the image, in a similar fashion to the spatial pyramid framework.

  The score for a Semantic Concept i on in a new image x is then

$$SC_i(x) = \sum_{k=1}^{N_i} w_k b_k(x) \quad (2)$$

  $SC_i(x)$ is the weighted sum of the scores on x of the individual SVMs, which we define base models b_k in the ensemble. The weights w_k are learned via cross-validation during training.

  For each video frame x, we concatenate all the models scores into a N-dimensional Semantic Model Vector, that is, a vector in which each dimension has a semantic meaning

$$SMV(x) = [SC_1(x), ..., SC_i(x), ..., SC_N(x)] \quad (3)$$

  We then use this concatenated vector as a regular feature for action modeling.

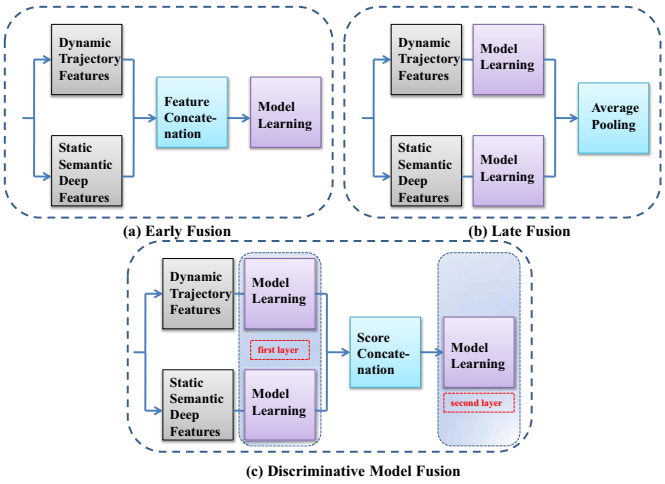

Figure 3: Illustration of heterogeneous feature fusion strategies.

In our experiments we built one separate vector for each of the two sets of classifiers.

3.2 Heterogeneous Information Fusion Strategies

Given the intuitive complementary nature of the different levels of semantic information we extract, it is natural to combine them into an integrated prediction system. There are many possible methods for combining heterogeneous information. As presented in Figure 3 we explored two standard ones (early fusion and average score based late fusion) and a discriminative one inspired by stacked SVMs approaches.

Early Fusion (EF): Early fusion strategy concatenates all the features into a single vector representation. The concatenated heterogeneous features are then directly fed into the multiclass SVM to train model for action categories.

Late Fusion (LF): In contrast to early fusion, where features are then combined into a universal representation, approaches for late fusion train models directly from each individual feature. The prediction scores from each model are then normalized to a common range and linearly combined in a late fusion step. In our experiments, we adopted the arithmetical mean as late fusion strategy.

Discriminative Model Fusion (DMF): we investigated the fusion performance following a two-layer model-learning strategy, instead of the one-layer model adopted in standard early and late fusion. In the first layer, as in the previous late fusion approach, models are trained individually on top of each descriptor. The predictions from the different models are then concatenated into a vector that is passed as input to the second layer, where an SVM with RBF kernel is learned on top of it. We did not tune any SVM parameter.

Discussion: Complementary information among heterogeneous features mentioned above may be of vital importance. Video-level feature description could be largely complementary to low-level dynamic visual features. For example, although two low-level features (HOF) quantized to the same fisher vector representation, videos may convey different semantics and false match could occur frequently (see Figure 2 for an illustration). Notwithstanding a significant difference

in visual appearance, the FV representation in the "PlayingPiano" clip closely resembles that of actions in "PlayingFlute".

4. EXPERIMENTS

In this section, we first describe our implementation details used in the experiments. Then we present recognition results on two popular datasets to examine the performance of the proposed approach. A comparison to the state-of-the-art methods is given at the end of this section.

4.1 Experimental Setting

We conduct experiments on two popular action datasets, namely HMDB51 [13] and UCF101 [12]. We summarize them and the experimental protocols as follows. The HMDB51 dataset consists 51 actions with 6,766 manually annotated clips which are extracted from a variety of sources ranging from digitized movies to Youtube. We follow the experimental settings in [13] and report the mean average accuracy over all classes. The UCF101 dataset [12] has been the largest action recognition dataset so far, and exhibits the highest diversity in terms of actions, with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions and so on. It contains 13,320 videos collected from YouTube and includes total number of 101 action classes. We perform evaluation on three train/test splits[1] and report the mean average accuracy over all classes.

Dense trajectories were extracted using the code provided by Wang et al.[11].

The extraction of deep features and CAFFE1K are conducted on a single Nvidia Telas K10 GPU with python interface, which speeds up by about 10 times than a decent Intel CPU with 8 cores.

The ConceptsWeb semantics were extracted using a CPU cluster running the LibSVM library [2].

In our experiments, we choose linear SVM as our classifier with the implementation of LIBLINEAR [5]. Then for multi-class classification, we use one-vs-all approach to perform action recognition and select the class with the highest score.

4.2 Results and Discussion

4.2.1 *Comparison with feature fusion*

We compute the performance of the different components of our system . To combine the representative capabilities, we investigate the performance of the fusion of dynamic trajectory features, static deep features as well as global taxonomy features. The combined features are used to train SVM model for each concept. There are three different fusion strategies–early fusion, late fusion and discriminative model fusion.

From Table 1 and Table 2, we could see that combining multiple types of dynamic trajectory descriptors can significantly improve recognition performance. Comparing across the alternative approaches, early fusion tends to generate better results with clear gains than late fusion. The reasons are two-folds. First, this should be partially ascribed to high-dimensional representations in early fusion we used. Second, the strategy of late fusion result in the heterogeneity among the confidence scores provided by different models. Such heterogeneity results from the variation of the discriminative capability of each model in a certain feature space, producing incomparable confidence scores at different numeric scales.

We also compare the performance of CAFFE features of the two fully-connected layers (fc_6 and fc_7), denoted as CAFFE-1K and CAFFE-4K. We observe from Table 2 that features from the six-layer are superior to those from the seventh layer. We speculate that the fully connected layer exert some negative effect on the features. Specifically, features extracted using "fc_6" (CAFFE-4K) obtains an mAP of 65.88% on UCF101 and 37.63% on HMDB51 dataset.

Then we augment the sematic features to the motion features. By including the static deep features, we can clearly see that for all methods, *i.e.*, "Dynamic + CAFFE4K", "Dynamic + CAFFE1K", the performance boost of feature fusion in both HMDB51 and UCF101 database is consistent and obvious, as compared and illustrated in Table 2. These results strongly prove that the contextual cues of static deep features are perfectly complementary to the low-level trajectory features. Besides, an intuitive analysis could also be easily provided. For instance, the action of "Basketball Dunk", we can know which object is contained (semantic static feature) and how is the object being played (motion feature). Notice that there are 37 classes out of 101 categories which are recognized with 100% accuracy. The full list UCF101 classes is reported in Table 4, where the categories for which we achieve perfect classification highlighted in blue.

Moreover, we also found that semantic ConceptsWeb features have less positive impact than static deep features on both evaluation datasets. For instance, on the UCF101 dataset, it yields 86.10% and 89.00% in mean average accuracy for ConceptsWeb and CAFFE1K features, respectively. It demonstrates that static deep features is more powerful in mitigating the semantic gap in comparison with intermediate-level semantic representation.

In Figure 4 we visualize the semantic meanings provided explicitly by the high level semantics features. For six actions, we sorted the weighs of the linear SVM trained on top of the CAFFE1K descriptor and picked the top three. Since each dimension (which is associated to a weight) in the CAFFE1K feature corresponds to one of the 1000 ImageNet categories, in this way we could extract the top 3 most discriminative concepts for each action. We then report the four keyframes from the entire dataset for which the concepts classifiers scored the highest. In this way we can qualitatively evaluate at the same time the correlation between action classes static concepts, and the quality of the indivisual concepts on the given dataset (which is not guaranteed since they were trained on different data). We can see that the most discriminative concepts for each action tend to make sense. For example "Biking" is mostly distinguished by *bicycle*, *mountain bike* and *moped*. It is interesting to notice that even in the cases in which the semantic meaning of a distinguishing concept does not appear to be correct (for example *abacus* for "Typing"), the classifier is actually recognizing useful information which is correlated to the action class (one or two hands on a device).

4.2.2 *Comparison with the state of the art*

We also compare our results with state-of-the-art methods on each dataset. Table 3 displays our best results and several recently published results in the literature. These method-

[1] http://crcv.ucf.edu/ICCV13-Action-Workshop/
[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm

Methods	Dynamic Trajectory Features				Static Deep Features	High Level Semantics	
	HOF	HOG	MBHx	MBHy	CAFFE-4K	CAFFE-1K	ConceptsWeb
UCF101	78.37%	73.65%	75.93%	77.91%	63.83%	65.88%	57.10%
HMDB51	49.19%	42.37%	41.11%	48.28%	37.63%	33.05%	25.23%

Table 1: Classification Performance comparison of individual descriptors: low-level dynamic dense trajectory features, intermediate-level static deep features and high-level semantic features.

Approaches	UCF101			HMDB51		
	EF	LF	DMF	EF	LF	DMF
HOF + HOG + MBHx + MBHy	87.37%	86.00%	88.00%	59.24%	58.06%	59.26%
CAFFE4K + CAFFE1K + ConceptsWeb	69.27%	68.27%	70.58%	40.24%	39.56%	41.27%
HOF + HOG + MBHx + MBHy + CAFFE4K	89.22%	88.48%	89.52%	62.68%	61.74%	62.81%
HOF + HOG + MBHx + MBHy + CAFFE1K	89.00%	88.80%	89.32%	61.00%	60.23%	61.74%
HOF + HOG + MBHx + MBHy + ConceptsWeb	86.10%	85.80%	86.20%	59.20%	58.23%	60.56%
HOF + HOG + MBHx + MBHy + Static	89.57%	89.53%	89.59%	62.71%	62.31%	62.88%

Table 2: Performance comparison of multiple fusion strategies with heterogeneous features.

s (e.g., mid-level parts [14]and deep architecture [10]) that utilize the responses of discriminative action parts combined with low-level features perform inferior to our approach with a certain margin consistently. For UCF-101 dataset, we achieve the best average accuracy 89.59%, which exceeds all the recent results reported in [3, 7, 8]. All these works are based on either low level trajectory features or ConvNets approach regardless of contextual semantic information. Note that extracting features in [3, 8] using neural networks needs more layers of neurons that incur significant number of additional parameters to be tuned, requiring much more training data and several weeks to train. While in many video classification tasks, the amount of available training data is far less from sufficient for training a neural network with too many layers. Moreover, since these adopted datasets are very challenging and have been widely used, it is interesting and competitive to obtain an absolute performance gain of 1.6% in comparison with [8].

For HMDB51 dataset, our result (62.88%) also outperforms the state-of-the-art approaches [7, 8]. To further prove the effectiveness of our approach on action recognition, we present the confusion matrix for HMDB51 dataset in Figure 5. By comparison, we observe that static deep features provide explicit enhancement to the low-level trajectory features. For instance, it could be easily found that the performance improvement on action of "kickball" is benefited from semantic description on objects such as "soccer", "soccer ball".

4.2.3 *Comparison with low-level and intermediate-level features*

We provide further a detail of the classification performance of state of the art static features (SIFT encoded with Fisher Vectors) and mid-level semantics such as Classemes and its derivatives PiCodes and MetaClasses.

The static information of Deep features and Semantics achieved good recognition performances and provided complementary information with respect to Dynamic descriptors. Therefore we analyzed alternative static descriptors, both *low-level* (i.e., SIFT [29]) and *intermediate-level* (i.e. semantic attributes [28]) on the UCF-101 dataset.

For low-level visual features, we extract SIFT and leverage the GMM model followed by Fisher Vector over the features extracted from all the frames in a video produce the whole clip-level representation. Specifically, we first densely extract local SIFT descriptors with a spatial stride of 4 pixels at 9 scales and the width of SIFT spatial bins is fixed as 8 pixels, which are the default settings in the VLFeat toolbox [30]. We learn a GMM dictionary sampled from a subset of 0.3 million SIFT descriptors. All descriptors are whitened after PCA processing to 64-dimension with a ratio of 0.5. We then conduct FV encoding and apply intra power normalization to the resulting super vectors. This state of the art *static* low-level feature obtains an average classification accuracy of 25.24% on the UCF-101 dataset, which as expected is significantly inferior to the performance of the *dynamic* low-level descriptors (Dense Trajectories). More interestingly, SIFT with FV encoding achieve recognition rates significantly lower than other static descriptors both at mid and high semantic levels.

Feature	Max	Average	Min
Classemes	53.16%	49.62%	-
Picodes	33.97%	52.39%	33.97%
Meta-Classes [28]	61.55%	45.76%	-
SIFT [29]+FV	25.24%		

Table 5: Performance comparison of multiple pooling strategies with intermediate-level semantic features. State of the art static descriptors (SIFT with Fisher Vectors encoding) provide a significantly worse performance.

Additionally, we compared the experimental performance of popular intermediate-level features (akin to semantic attributes) which were recently introduced by Bergamo et al. [28]: Classemes and its derivatives Picodes and Meta-Classes. Since each of those descriptors are extracted from individual frames, we tested several basic pooling approaches (Max, Average, Min) in order to produce a video-level feature representation which aggregates the frames responses. As reported in Table 5, we observe that such intermediate-level features achieved less competitive performance (best average accuracy of 61.55% over all feature types and aggregation strategies) than their Deep counterparts at the same semantic level (Deep features, 63.83%).

5. CONCLUSION

We proposed a novel unified framework that jointly combines dynamic trajectory features and exploits the class re-

Methods	Ours	[17]	[8]	[11]	[3]	[12]	[7]	[18]	[10]	[14]	[19]	[32]	[21]
UCF101	**89.6%**	84.2%	88.0%	85.9%	65.4%	43.9%	87.9%	83.5%	87.7%	-	73.1%	-	-
HMDB51	**62.9%**	56.3%	59.4%	57.2%	-	-	61.1%	55.9%	59.8%	37.2%	49.9%	48.7%	66.8%

Table 3: Performance comparison with state-of-the-art methods.

YoYo	Archery	JumpRope	Basketball	HeadMassage	FrisbeeCatch	ApplyLipstick	BoxingSpeedBag	*BoxingPunchingBag*
Punch	Bowling	Kayaking	*BenchPress*	*HorseRiding*	JavelinThrow	BaseballPitch	CricketBowling	*TrampolineJumping*
Swing	*Fencing*	*Knitting*	*FrontCrawl*	*JumpingJack*	MoppingFloor	BrushingTeeth	MilitaryParade	*VolleyballSpiking*
Biking	Haircut	LongJump	IceDancing	PlayingDhol	*ParallelBars*	JugglingBalls	SoccerJuggling	FieldHockeyPenalty
Diving	PullUps	*Billiards*	PlayingDaf	*PommelHorse*	PizzaTossing	PlayingGuitar	WalkingWithDog	*RockClimbingIndoor*
Lunges	PushUps	GolfSwing	*StillRings*	TennisSwing	*PlayingCello*	PlayingViolin	*WritingOnBoard*	
Mixing	Rafting	Hammering	UnevenBars	ThrowDiscus	PlayingFlute	SkateBoarding	*FloorGymnastics*	
Rowing	Shotput	*HorseRace*	BalanceBeam	WallPushups	*PlayingPiano*	*SoccerPenalty*	TableTennisShot	
Skiing	Surfing	Nunchucks	BlowDryHair	BabyCrawling	*PlayingSitar*	SumoWrestling	*BodyWeightSquats*	
Skijet	Drumming	*PoleVault*	CliffDiving	BandMarching	*PlayingTabla*	ApplyEyeMakeup	CuttingInKitchen	
TaiChi	HighJump	*SalsaSpin*	CricketShot	*BreastStroke*	RopeClimbing	BasketballDunk	*HandstandPushups*	
Typing	HulaHoop	SkyDiving	*HammerThrow*	*CleanAndJerk*	ShavingBeard	*BlowingCandles*	HandstandWalking	

Table 4: Illustration of 37 categories (in Blue) which obtained 100% accuracy in UCF101 dataset.

lationship via static deep features and high-level semantic descriptors for improving action classification performance. This unique capacity distinguishes the proposed method from most of the existing works that often adopted low-level dense features without considering the inter-class semantic correlations. Our investigation also implies that static deep and semantic features are largely complementary to low-level dynamic trajectory features. The SVM based fusion approach provided the best framework to combine the information coming from the heterogeneous static and dynamic features at different semantic levels. Extensive empirical evaluations proved the effectiveness of the proposed framework, which achieves an accuracy of 89.59% and 62.88% on the well-known UCF-101 and HMDB51 benchmarks, respectively, which compare favorably with the state-of-the-arts.

We plan to explore multiple classification kernels with heterogenous types of features. We also plan on learning the temporal evolution of semantic descriptors representing more complex activities.

6. ACKNOWLEDGEMENTS

We would like to thank Liangliang Cao, Noel Codella, John Smith, Quoc-Bao Nguyen and Quanfu Fan for fruitful discussions. This work was supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057 and Faculty Research Award by NEC Laboratories of America, Inc. This work was supported in part by National Science Foundation of China (NSFC) 61429201.

7. REFERENCES

[1] M. Merler, B. Huang, L. Xie, G. Hua and A. Natsev. Semantic model vectors for complex video event recognition. In *IEEE Transactions on Multimedia*, 2013.

[2] Y. Jia. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. *http://caffe.berkeleyvision.org/*, 2013.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014.

[4] A. Krizhevsky, I. Sutskever and G.E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of Advanced in Neural Inforamtion Processing System*, 2012.

[5] R. Fan, K. Chang, C. Hsieh, X. Wang and C. Lin. LIBLINEAR: a library for large linear classification. In *Journal of Machine Learning Research*, 2008.

[6] H. Wang, C. Yuan, W. Hu, H. Hu and C. Sun. Action recognition using nonnegative action component representation and sparse basis selection. In *Transaction on Image Processing*, 2014.

[7] X. Peng, L. Wang, X. Wang and Y. Qiao. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. In *arXiv*, 2014.

[8] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of Advanced in Neural Inforamtion Processing System*, 2014.

[9] Large scale visual recogntion challenge. *http://www.image-net.org/challenges/LSVRC/2012/*, 2012.

[10] X. Peng, L. Wang, Y. Qiao and Q. Peng. Boosting VLAD with Supervised Dictionary Learning and Higher-Order Statics. In *Proceedings of the European Conference on Computer Vision*, 2014.

[11] H. Wang and C. Schmid. Action recogntion with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[12] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *arXiv:1212.0402*, 2012.

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[14] M. Sapienza, F. Cuzzolin and P. S. Torr. Learning discriminative space-time action parts from weakly labelled videos. In *International Journal of Computer Vision*, 2014.

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

[16] I. Laptev. On space-time interest points. In *International Journal of Computer Vision*, 2005.

[17] J. Wu, Y. Zhang, W. Lin. Towards good practice for action video encoding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014.

[18] Z. Cai, L. Wang, X. Peng, Y. Qiao. Multi-view super vector for action recogntion. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014.

[19] O. V. R. Murthy and R. Goecke. Ordered Trajectories for Large Scale Human Action Recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013.

[20] Z. Ma, Y. Yang, N. Sebe, K. Zheng and A. G. Hauptman. Multimedia event detection using a classifier-specific intermediate representation. In *IEEE Transactions on Multimedia*, 2013.

[21] X. Peng, C. Zhou, Y. Qiao and Q. Peng. Action recognition with stacked fisher vectors. In *Proceedings of the European Conference on Computer Vision*, 2014.

[22] J. Liu, Y. Yang and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2009.

[23] J. Liu, B. Kuipers and S. Savarese. Recognizing Human Actions by Attributes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2011.

Figure 4: Top three discriminative concepts (out of 1,000 ImageNet categories) for six action classes.

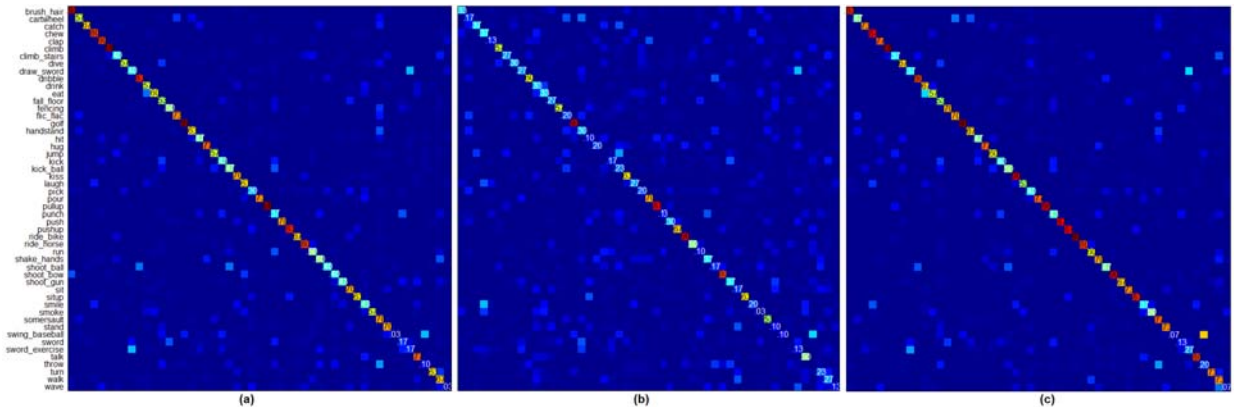

Figure 5: Performance comparison in terms of confusion matrix on the HMDB-51 dataset. (a) Dynamic Trajectory Feature. (b) Combination of Static Deep and Semantic Features. (c) Fusion of Dynamic and Static Feature

[24] B. Yao, X. Jiang, A. Khosla and L. Feifei. Human action recogntion by learning bases of acion attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[25] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed and N. Jaitly. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. In *IEEE Signal Processing Magazine*, 2012.

[26] Y. Lecun, K. Kavukcuoglu and C. Farabet. Convolutional networks and applications in vision. In *Proceedings of IEEE International Symposium on Circuits and Systems*, 2010.

[27] O. Russakovsky, J. Deng, H. Su and A. Berg. ImageNet large scale visual recogntion challenge. In *arXiv*, 2014.

[28] A. Bergamo and L. Torresani. Classemes and Other Classifier-based Features for Efficient Object Categorization. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2013.

[29] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, 2004.

[30] A. Vedaldi and B. Fulkeson. VLFeat: An open and portable library of computer vision algorithms. 2008.

[31] J. Chen, Y. Cui, G. Ye, D. Liu and S.-F. Chang. Event-Driven semantic concept discovery by exploiting weakly tagged internet images. In *Proceedings of the IEEE Conference on Multimedia and Retrieval*, 2014.

[32] Z.-Z. Lan, L. Bao, S. Yu. Multimedia classification and event detection using double fusion. In *Multimedia Tools and Applications*, 2014.

[33] S. Sadanand and J. Corso. Action bank: a high-level representation of activity in video. In *CVPR*, 2012.

[34] D. Oneata, J. Verbeek and C. Schmid. The Lear submission at Thumos 2014. In *THUMOS Workshop*, 2014.

[35] R. Arandjelovic and A. Zisserman. All about VLAD. In *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013.

[36] Y.-G. Jiang, J. Liu and A. Zamir, G. Toderici,I. Laptev, M. Shah and B. Sukthankar. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/, 2014.

[37] J. Cai, Z. Zha, M. Wang, S. Zhang and Q. Tian An Attribute-assisted Reranking Model for Web Image Retrieval. In *Transaction on Image Processing*, 2014.