Harvesting Motion Patterns in Still Images from the Internet

Jiajun Wu (jiajunwu.cs@gmail.com)
Yining Wang (ynwang.yining@gmail.com)
Zhulin Li (li-zl12@mails.tsinghua.edu.cn)
ITCS, Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084

Zhuowen Tu (ztu@ucsd.edu)
Department of Cognitive Science, University of California, San Diego, La Jolla, 92093

Abstract

Most vision research on motion analysis focuses on learning human actions from video clips. In this paper, we investigate the use of still images, rather than videos, for motion recognition. We present evidence from both human cognition and computer vision that still images do indeed contain a wealth of information about motion patterns. Our contributions are three-fold. First, we automatically determine classes of motions that can effectively be characterized by still images. To make this determination we introduce the notions of motion verbs (M-verbs) and motion phrases (M-phrases); these refer to linguistic concepts motivated by visual cognition and are not restricted only to motions performed by humans. Second, we build UCSD-1024, a large dataset distilled from more than two million still images. These images come from 1,024 categories of motion; we use crowdsourcing to provide human validation of the motion categories. Third, we exploit motion patterns from UCSD-1024 using a weakly-supervised learning strategy and demonstrate performance competitive with state-of-the-art computer vision action classification methods.

Keywords: motion pattern discovery; image with implied motion; visual perception

Introduction

Action recognition has long been a topic of interest in vision research. In addition to traditional computer vision methods that aim to learn effective action models from videos (Sadanand & Corso, 2012; Soomro, Zamir, & Shah, 2012), much recent research has investigated the use of still images (Delaitre, Laptev, & Sivic, 2010; Khan et al., 2013). While certainly inspiring, we note that these recent still-image-based approaches ignore two basic questions underlying the use of still images in action recognition: Given that actions are intrinsically dynamic processes, are still images rich enough for action recognition of any kind? If so, what types of actions can be captured by still images?

In this paper, we demonstrate that still images are indeed rich enough for use in motion understanding; in the process, we also characterize those motions that are recognizable even in still images. This is an interdisciplinary topic at the intersection of cognitive science, computer vision, natural language processing, and linguistics. We first discuss findings in cognitive science that provide theoretical support of our claim of still image richness. We then use the machine learning strategy of multiple instance learning (MIL) to further demonstrate the expressiveness of still images by learning from still frames in videos. We propose motion verbs (M-verbs) and motion phrases (M-phrases) for the class of verbs and phrases describing motions that can be effectively conveyed by still images. In linguistics, it is known that verbs can be divided into four categories: states, activities, achievements, and accomplishments, based on their telicity and continuity (Rothstein, 2004). After we introduce these linguistic categories, we then discuss the relationship between our notion of M-verbs and the linguistic categorization of verbs before drawing further conclusions on the linkage between M-verbs and continuity of verbs. As shown in Figure 2, these findings guide us to a novel setting bridging verbs in linguistics and motions in computer vision.

With a foundation of M-verbs and M-phrases, we then build the large UCSD-1024 dataset from more than two million images across 1,024 categories. As our first step to UCSD-1024, we develop a semi-supervised knowledge expansion framework to determine precisely which phrases (selected from a large corpus and a small number of labeled seeds) correspond to actions that can be effectively conveyed by still images. Combining this semi-supervised output with corroboration from Amazon Mechanical Turk, we construct a dictionary of 1,024 M-phrases. We subsequently use this dictionary, in concert with the Google and Bing image search engines, to build UCSD-1024.

Learning mid-level representations is a popular topic in computer vision (Lim, Zitnick, & Dollár, 2013; Q. Li, Wu, & Tu, 2013). Here we learn a dictionary for motions using a hierarchical model based on mid-level representations on an eighty motion subset of UCSD-1024 that we henceforth refer to as UCSD-80. We then perform action classification on Stanford 40 (Yao et al., 2011) and obtain encouraging results.

Related Work

Most existing action recognition research in computer vision is based on video clips (Sadanand & Corso, 2012; Soomro et al., 2012). Recently, researchers have pursued action recognition in still images (Delaitre et al., 2010; Khan et al., 2013). Delaitre et al. (2010) performed action recognition in still images using a combination of bag-of-feature methods and part-based representations, and built a dataset of seven categories and 968 Flickr images. Yao and Fei-Fei (2010) proposed Grouplet, a structural representation for interactions between humans and objects, and the PPMI dataset of seven types of activities. However, none of these explicitly address the underlying questions for action recognition: how expressive are still images and what types of actions can still images effectively convey. Finally, the largest existing dataset is Stanford 40 (Yao et al., 2011), comprised of 9,532 images in 40 categories. By comparison, the proposed UCSD-1024 dataset is distilled from roughly two million images across 1,024 categories. Moreover, UCSD-1024 is not restricted only to motions performed by humans.

In linguistics, verb categorization dates back to Aristotle's trichotomy (Taylor, 1977); see the representative works (Tenny, 1987; Rothstein, 2004) for further discussion. More recently, (Taylor, 1977) linked continuity and tense and (Fleck, 1996) discussed the spatial and temporal properties of verbs topologically.

The task of expanding M-phrases is related to the *thesaurus extraction* task insofar as both aim to create a list of terms. The use of online texts for thesaurus extraction was first investigated in (Jannink, 1999). Curran and Moens (2002) proposed an automatic thesaurus extraction algorithm based on syntactic structures and word distributions of online texts. The *C-value/NC-value* method proposed in (Frantzi, Ananiadou, & Mima, 2000) uses both syntactic features and statistical measures for phrase extraction. Our task differs from thesaurus extraction in that the definition of M-phrases involves both semantic and visual understanding of the phrases, making the problem much harder.

Motions in Still Images

In this section, we discuss motion recognition in still images from a cognitive science perspective. We then use a standard machine learning method, multiple instance learning (MIL), to demonstrate the potential effectiveness of still images in motion recognition.

A Cognitive Science View

Still images with implied motion have long been of interest in psychology and cognitive science. Freyd (1983) showed that visual stimuli in which motion is only implied, such as frozen motion photographs, could nonetheless prompt the brain to rapidly and automatically extrapolate motion paths. Kourtzi and Kanwisher (2000) subsequently demonstrated that viewing static images with implied motion could prompt activity in medial temporal/medial superior temporal cortex (MT/MST). Proverbio, Riva, and Zani (2009) showed that such observations could also enhance the activity of movement-related brain areas. Thus, these demonstrations provide strong neurological evidence that motion can be understood even in still images.

Recent work from Boroditsky's group (Winawer, Huk, & Boroditsky, 2010, 2008; Dils & Boroditsky, 2010) provides a more thorough theoretical cognitive foundation. Winawer et al. (2010) connects visual imagery of motion with perceptual motion, (Winawer et al., 2008) relates still images of actions with human cognition, and (Dils & Boroditsky, 2010) links visual motion understanding with motion language. These studies inspire us to integrate cognition and linguistics into vision applications.

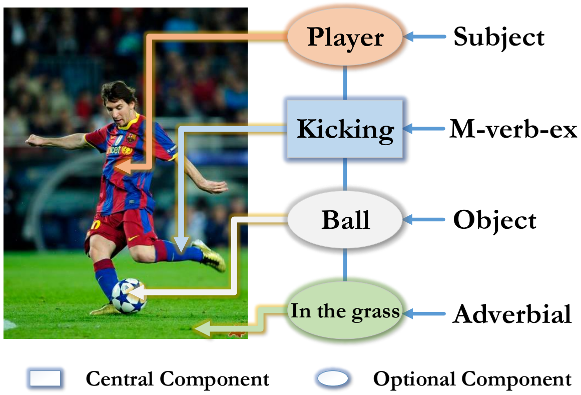


Figure 1: An illustration of M-phrases

	Stages	Telic
States	-	-
Activities	+	-
Achievements	-	+
Accomplishments	+	+

Table 1: Types of verbs from linguistics (Rothstein, 2004)

A Demonstration with Multi-Instance Learning

Here we show that still images contain rich information about motion patterns. We apply multiple instance learning (MIL) (Andrews, Tsochantaridis, & Hofmann, 2002) to video categories to discover the most relevant frame and then learn models for the corresponding action in that frame using video sequences as bags and frames as instances.

Specifically, we first randomly select seven categories from HMDB-51 (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011); within each category we then randomly select five video sequences. We next compute GIST (Oliva & Torralba, 2006) for a number of frames. Treating each video sequence as a positive bag and a collection of irrelevant images as a negative bag, we learn one instance-level classifier for each category using mi-SVM (Andrews et al., 2002). We subsequently sample a few more video sequences in each category and perform video classification on them using average-voting on images with the learned classifiers. The classification accuracy is 64.7%, providing strong evidence that still images can effectively convey motions.

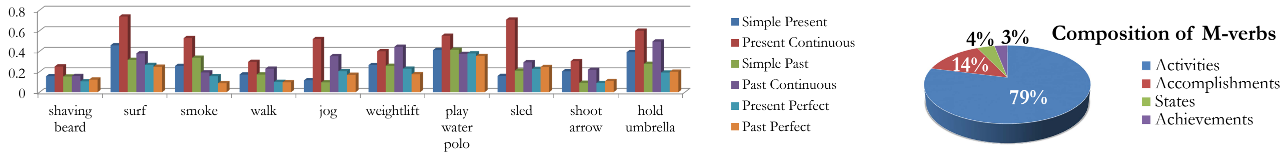
Verbs at a Glance

In this section, we pursue a new approach: in order to characterize motions/verbs that can be effectively captured in still images, we must first introduce a novel partition of verbs. We provide notions meant to refer to those verbs or phrases that describe motions that can be effectively conveyed by still images. In particular, these still-image-conveyable motion verbs and phrases are called M-verbs and M-phrases, respectively.

M-verbs and M-phrases

We define *motion verbs (M-verbs)* as verbs that can be effectively conveyed by still images. We roughly categorize M-verbs into three classes:

- Simple verbs: run, laugh, swim,



(a) Percentage of images labeled as positive ones by all three workers with respect to ten categories and six tenses. Note that present continuous tense is almost consistently the highest.

(b) The composition of M-verbs with respect to four verbal categories.

Figure 2: M-verbs and M-phrases with respect to tenses and verbal categories.

- Compound verbs: cliff diving, ice skating,
- Special verbal phrases: push up, pull up.

Noting that there are verbs that cannot be associated with a visual impression in the absence of sufficient contextual information, *e.g.* closing v.s. closing eyes, we define *extended motion verbs (M-verbs-ex)* as a superset of M-verbs containing all verbs that could potentially convey visual motion in an appropriate context. Based on M-verbs-ex, we propose *motion phrases (M-phrases)*, which extend M-verbs by incorporating subjects, objects, and adverbials. As shown in Figure 1, an M-phrase contains an M-verb-ex as its central component with subject, object, and adverbial as optional components.

M-phrases are both flexible and expressive. The use of subjects and objects makes them capable of describing possible motions in still images. Note that here we do not require each image to contain an entire human body. For instance, M-phrases like *dog running*, *horse galloping*, *clapping hands*, and *scowling* describe motions that could be accomplished with non-human creatures or parts of human bodies.

M-verbs from a Linguistic Point of View

Aristotle's trichotomy classified verbs into three categories: state-verbs, energeia-verbs, and kinesis-verbs. Later linguists (Taylor, 1977; Tenny, 1987; Fleck, 1996) further developed the partition and it is now commonly agreed that a verb or verbal constituent belongs to one of four categories:

- States: is (hirsute), love (a school),
- Activities: swim, talk,
- Achievements: discover (America), pass (an exam),
- Accomplishments: build (a house), stab (Caesar).

These four categories are also associated with two crucial aspectual properties: whether the verbs in question can appear in progressive forms (Stages) and whether they occur with movement towards an endpoint (Telic) (Rothstein, 2004).

M-verbs and M-phrases are sets of verbal constituents describing motions that can be effectively conveyed by still images. This focus thus brings us to a novel setting for which existing theories in linguistics and vision fail to fit as closely as necessary. Here we would like to explore some possible connections between motions in vision and verbs in linguistics that will help us better fit our focus.

We select 100 M-verbs from 287 seeds (introduced in the next section). We then manually classify these verbs into states, activities, achievements, and accomplishments. To ensure accuracy, we further ask three individuals to independently label each verb; any verbs for which these labels disagree are given to an expert linguist for a final label. We see from Figure 2b that most M-verbs are either activities or accomplishments. This, together with the linguistic knowledge from Table 1, indicates that continuity (stage) plays a key role in the composition of M-verbs. In this sense, M-verbs connect concepts in linguistics and vision.

Topology, Continuity, and Tense

As mentioned previously, there is a strong tie between continuity and motion verbs. In linguistics, boundaries of space and time of verbs can be modeled from a view in topology (Fleck, 1996), which is also closely related to the use of different tenses with respect to the continuity of verbs.

We are thus motivated to explore how the use of different tenses might help our construction of M-verbs and M-phrases in UCSD-1024. We employ Amazon Mechanical Turk for corroboration. Specifically, we select ten M-verbs and six tenses. For each phrase as a combination of an M-verb and a tense, we crawl 1,000 images from Google image search using that phrase as query. We then ask workers to classify whether each crawled image is relevant for that query.

The result is shown in Figure 2a. We see that present continuous tense yields the highest intra-annotator agreement in nine of the ten categories. These empirical statistics (and the linguistic analysis considered earlier) lead us to use present continuous tense in all M-verbs and M-phrases.

Building UCSD-1024

We now introduce UCSD-1024. We aim to exploit the rich knowledge in still images by building the largest still image motion database with M-phrases. UCSD-1024 is distilled from over two million images across 1,024 categories.

Collecting Seeds

We first collect M-phrases to serve as seeds for dictionary expansion. We collect 101 M-phrases from UCF101 (Soomro et al., 2012), 40 M-phrases from the Stanford 40 Action Dataset (Yao et al., 2011), 90 M-phrases for common sports, and 11 M-phrases for common facial expressions. After removing duplicates and merging M-phrases with similar meanings, these sources contribute 237 M-phrases.

We obtain additional seeds by exploiting the inherent structure of M-phrases. We employ 13 common subjects, 78 M-verbs, 15 objects, and 12 adverbials to generate ca. $200,000$ M-phrase candidates. We then determine the popularity of

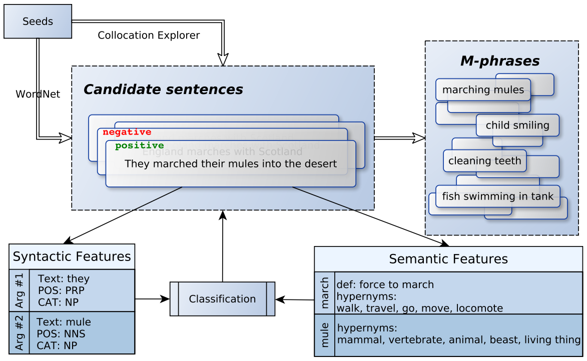


Figure 3: Framework for M-phrases expansion

these candidates using search engines; by manually selecting 50 M-phrases from the 200 most popular, we arrive at $237+50=287$ seed M-phrases.

Expanding M-phrases

From these 287 seeds, we expand our dictionary of M-phrases via a three-step framework shown in Figure 3. We first use seed verbs (M-verbs included in our seed dictionary, such as *throw*, *ride*) to crawl sentences from the Internet. We then extract syntactic and semantic features based on crawled sentences and apply supervised classification to determine whether a sentence contains M-phrases. Finally, after classifying the sentences, we design a rule-based extractor to extract M-phrases from each containing sentence.

Data source: We crawl sentences from *Collocation Explorer*, a system that automatically detects collocations from the British National Corpus. The primary advantage of *Collocation Explorer* is its ability to return sentences containing user-specified verbs. By using our seed M-verbs, we obtain sentences of higher quality than randomly picking sentences from a corpus.

In addition, *Collocation Explorer* allows advanced search patterns, further contributing to the precision of our system. We use our seeds to crawl 45,140 sentences.

Syntactic and semantic features: We design both syntactic and semantic features for our unsupervised learning framework. Features used include syntactic categories and head-noun Part-Of-Speech (POS) tags.

To address polysemy, we use word sense disambiguationto assign each word in a sentence a corresponding "synset" in WordNet (Fellbaum, 2010) that represents the meaning of the word. We then use bag-of-word features on the synset definitions to separate words with different meanings.

Supervised classification and M-phrases extraction: We used support vector machines (SVMs) to classify each sentence as either "containing" or "not containing" each M-phrase type. We randomly picked 1,024 sentences as training data and manually labeled them. We determine the SVM parameters via five-fold cross-validation.

Finally, we extract M-phrases from each containing sentence. We being by anchoring the key M-verb in each sentence, and then using the Enju parser to locate the arguments of the verb. Only the head nouns of the arguments will appear in the final M-phrases and some post-processing steps such as changing all verbs into their -ing forms were carried out to make the M-phrases suitable for image collection.

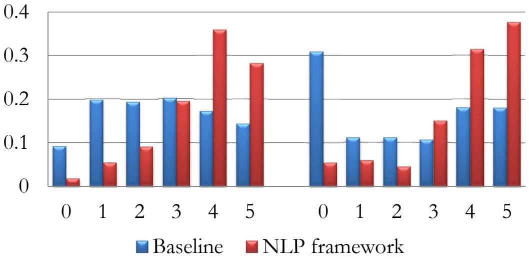


Figure 4: Percentage of images with different ratings, annotated by two specialists

Quantitative evaluation: We evaluate the quality of our M-phrase expansion algorithm by rating the images crawled. Specifically, we randomly crawled 500 images using M-phrases we expanded and another 500 images using phrases generated by a baseline approach: randomly combining seed nouns and seed verbs. We then ask annotators to rate the 1,000 images without revealing to them which method we used for a particular image. The ratings are shown in Figure 4, where images score "5" when most relevant to motions and score "0" when least relevant. Our M-phrase expansion results in more consistent images, thereby validating our M-phrase expansion algorithm.

Enriching M-phrases Using Crowdsourcing

Both the British National Corpus and WordNet are "closed universes" — they do not include every valid M-phrase. To open the "closed universe", we recruited Amazon Mechanical Turk (AMT) workers to provide additional phrases. We asked each user to provide 10 phrases that could be effectively conveyed by still images. Each answer was required to be five words or fewer and to not make a complete sentence. To ensure diversity, we limited each user to at most 50 phrases. We manually rewrote 70 phrases most suitable to our task as M-phrases and added them to our dictionary.

From the combined output of our expansion framework and our AMT task, we selected 1,024 M-phrases. We then built the UCSD-1024 image dataset with the help of Internet image search engines and crowdsourcing. We first used Google and Bing to crawl 1,000 images for each M-phrase. When submitting queries, we restricted the results to be photos only. We then removed all broken links, any images smaller than 100×100, and duplicates. For each category, we retained 1,000∼1,900 images for further processing.

Images provided by search engines are diverse, but also noisy. We used AMT to recruit workers to assist in cleaning up the data. For this task, we created a large number of hits, each of which contained 40 Internet images crawled with one M-phrase. Workers were asked to decide whether each of the images was relevant to the M-phrase query, with each hit assigned to three different workers.

M-verb		S + M-verb-ex		M-verb-ex + O / Ad		S + M-verb-ex + O / Ad
crawling	throwing	whistle blowing	child running	raising hands	applying eye makeup	wind blowing leaf
marching	applauding	water flowing	man smoking	pushing against wall	applying lipstick	boat drifting on water
brushing	diving	leaf swirling	fish swimming	delivering ball	blowing dry hair	fish swimming in tank
pushing	walking	cat running	kid skiing	brushing hair	blowing bubbles	feather drifting past window
cycling	smiling	dog barking	woman smoking	cooking dinner	blowing candles	dog licking hand
jogging	dancing	man sailing	baby crawling	lifting box	brushing teeth	bird clapping wing
archering	dunking	military marching	band playing	climbing rock	cutting trees	face being angry
bowling	drinking	car running	baby wailing	closing eyes	raising eyebrows	face being disgusted
boxing	fishing	child clinging	child writing	fixing car	fixing bike	face being surprised
kayaking	bathing	dog baying	dog eating	playing badminton	playing football	dentist cleaning tooth
coughing	decanting	fish swimming	girl dancing	playing guitar	playing cello	people crowding street
dabbling	harvesting	girl walking	girl pouting	ascending mountain	assembling car	parent protecting child
refueling	spinning	man leaping	man sitting	bonding with child	brushing wall	pitcher delivering ball
spitting	telephoning	potato sprouting	train derailing	cheering child	conditioning hair	squirrel leaping from tree
undressing	yelling	tree swaying	water bubbling	cleaning fingernail	cleaning stove	smoke rising from fire
dancing	crawling	water pouring	woman biting	disciplining child	drinking soda	spider spinning web

Table 2: A subset of M-verbs and M-phrases we employ

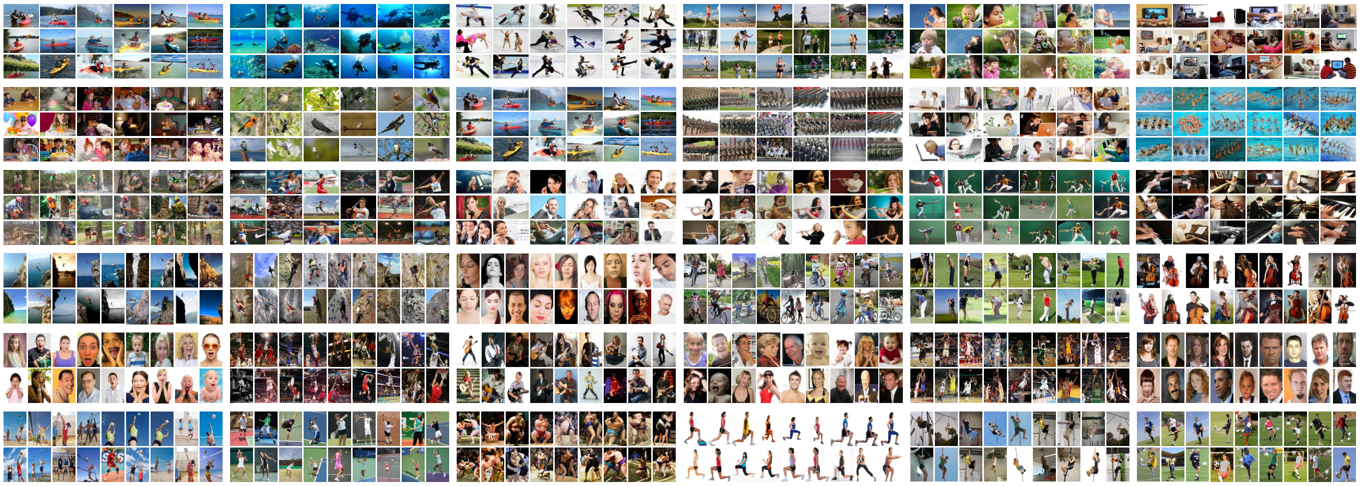


Figure 5: A subset of the UCSD-1024

We vet submission quality as follows: among the three submitted labels for an image, we regard one of them as *questionable* if it differs from the other two. For each hit, if over 50% of its 40 submitted labels are questionable, we consider it a *bad* hit. For each worker, if his bad hits compose over 30% of all his submissions, we reject all his submissions, block him from further participation, and reassign other workers to finish his hits. For other workers, we reject only "bad" submissions.

The final submissions are highly consistent. Quantitatively, 30.1% of the images are labeled as positive by at least two of the three workers, and 46.2% are labeled as positive by all three workers. For higher accuracy, we retain only images labeled as positive by all three labelers. Figure 5 shows a subset of UCSD-1024 after quality control.

Human Validation of UCSD-1024

In this section, we discuss the data consistency of UCSD-1024 and verify that UCSD-1024 keeps both intra-category consistency and annotator agreement.

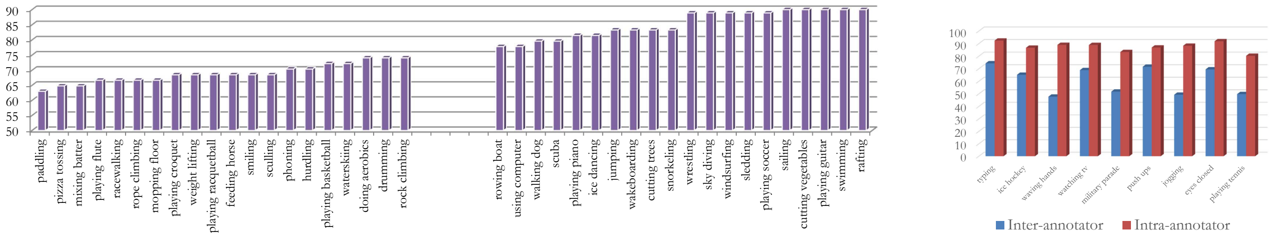
For intra-category consistency, we conduct another AMT experiment: We randomly select two images from one category, and eight images from the others (so that no two of the eight are from the same category). Without revealing the labels we then ask each user to pick out two images that they think belong to the same category from the ten images. We repeat the experiment ten times for each category. The results show that the percentage of correct submissions varies from 63.0% to 96.3% for different categories. Figure 6 demonstrates the twenty M-phrases with highest or lowest agreement and provides examples. The high percentage shows the intra-category agreement of images obtained from the Internet and AMT.

Apart from intra-category consistency, we also consider the intra- and inter-annotator agreement of UCSD-1024.

Intra-annotator agreement (InAA) measures the consistency of annotation by the same annotator. We invited 20 annotators to annotate the same set of images twice, with an one-month gap in between, and investigate the consistency of the labels. Inter-annotator agreement (ItAA) can be conveniently measured via AMT, where each image was labeled by three different workers. Although we discarded all images with inconsistent labels, the percentage of consistently labeled images still provides useful information on the consistency of UCSD-1024. Figure 6b shows the intra- and inter-annotator agreement for several motion categories.

Learning Hierarchical Action Models

To demonstrate the richness of UCSD-1024, we further develop a hierarchical model for action classification. For experimental use, we employ UCSD-80, a subset of UCSD-1024, for more efficiency at the cost of some expressiveness.



(a) 20 categories with lowest and highest intra-category consistency

(b) Intra-/inter-annotator agreement

Figure 6: From (a) we can see that for categories with relatively low consistency, the images within each category are still highly consistent. The low score is largely due to human variance in classifying images between similar classes.

Method	Accuracy	# Features
HOG+LBP	23.4%	2400×21
Visual Concepts	27.5%	716×21
Motion (ours)	29.6%	80×21
Motion+VC+HOG (ours)	**33.1%**	80×21+716+2400
Object Bank	32.5%	—

Table 3: Comparison of feature accuracy and length for different action classification approaches applied to Stanford 40.

Following (Q. Li et al., 2013), we learn 716 mid-level classifiers using Visual Concepts (Q. Li et al., 2013). We then train one-vs-all SVMs for the 80 categories on the response of the mid-level classifiers, resulting in 80 layered models for motions.

We test these models in action classification with Spatial Pyramid Matching (Lazebnik, Schmid, & Ponce, 2006) on Stanford 40 (Yao et al., 2011); the number of features are therefore multiplied by $1+2\times 2+4\times 4=21$. Table 3 shows that our action model achieves better results with a much smaller size of dictionary than Visual Concepts (Q. Li et al., 2013). Combining features from lower layers, our method, with only very light supervision in M-phrases expansion, outperforms Object Bank (L.-J. Li, Su, Fei-Fei, & Xing, 2010) which requires fully supervised bounding boxes in training.

Acknowledgments

This work is supported by NSF IIS-1216528 (IIS-1360566) and NSF CAREER award IIS-0844566 (IIS-1360568). We thank Patrick Gallagher for his helpful suggestions.

References

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2002). Support vector machines for multiple-instance learning. *Advances in Neural Information Processing Systems*.

Curran, J., & Moens, M. (2002). Improvements in automatic thesaurus extraction. *ACL workshop on Unsupervised lexical acquisition*.

Delaitre, V., Laptev, I., & Sivic, J. (2010). Recognizing human actions in still images: a study of bag-of-features and part-based representations. *British Machine Vision Conference*.

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences*, *107*(37), 16396–16400.

Fellbaum, C. (2010). *Wordnet*. Springer Netherlands.

Fleck, M. M. (1996). The topology of boundaries. *Artificial Intelligence*, *80*(1), 1–26.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, *3*(2), 115–130.

Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. *Perception & Psychophysics*, *33*(6), 575–581.

Jannink, J. (1999). Thesaurus entry extraction from an online dictionary. *Proceedings of Fusion*.

Khan, F. S., Anwer, R. M., van de Weijer, J., Bagdanov, A. D., Lopez, A. M., & Felsberg, M. (2013). Coloring action recognition in still images. *International Journal of Computer Vision*, 1–17.

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human mt/mst by static images with implied motion. *Journal of Cognitive Neuroscience*, *12*(1), 48–55.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: a large video database for human motion recognition. *IEEE International Conference on Computer Vision*.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Conference on Computer Vision and Pattern Recognition*, *2*, 2169–2178.

Li, L.-J., Su, H., Fei-Fei, L., & Xing, E. P. (2010). Object bank: A high-level image representation for scene classification & semantic feature sparsification. *Advances in Neural Information Processing Systems*.

Li, Q., Wu, J., & Tu, Z. (2013). Harvesting mid-level visual concepts from large-scale internet images. *IEEE Conference on Computer Vision and Pattern Recognition*.

Lim, J. J., Zitnick, C. L., & Dollár, P. (2013). Sketch tokens: A learned mid-level representation for contour and object detection. *IEEE Conference on Computer Vision and Pattern Recognition*.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, *155*, 23–36.

Proverbio, A. M., Riva, F., & Zani, A. (2009). Observation of static pictures of dynamic actions enhances the activity of movement-related brain areas. *PLoS One*, *4*(5), e5389.

Rothstein, S. (2004). Verb classes and aspectual classification. *Structuring Events: A Study in the Semantics of Lexical Aspect*, 1–35.

Sadanand, S., & Corso, J. J. (2012). Action bank: A high-level representation of activity in video. *IEEE Conference on Computer Vision and Pattern Recognition*.

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *CRCV-Tech Report-12-01*.

Taylor, B. (1977). Tense and continuity. *Linguistics and philosophy*, *1*(2), 199–220.

Tenny, C. L. (1987). Grammaticalizing aspect and affectedness. *Ph.D. Thesis*.

Winawer, J., Huk, A. C., & Boroditsky, L. (2008). A motion aftereffect from still photographs depicting motion. *Psychological Science*, *19*(3), 276–283.

Winawer, J., Huk, A. C., & Boroditsky, L. (2010). A motion aftereffect from visual imagery of motion. *Cognition*, *114*(2), 276–284.

Yao, B., & Fei-Fei, L. (2010). Grouplet: A structured image representation for recognizing human and object interactions. *IEEE Conference on Computer Vision and Pattern Recognition*.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., & Fei-Fei, L. (2011). Human action recognition by learning bases of action attributes and parts. *IEEE International Conference on Computer Vision*.