Atomic Action Features: A New Feature for Action Recognition

Qiang Zhou[1] and Gang Wang[1,2]

[1] Advanced Digital Sciences Center, Singapore
Zhou.Qiang@adsc.com.sg
[2] Nanyang Technological University, Singapore
wanggang@ntu.edu.sg

Abstract. We introduce an atomic action based features and demonstrate that it consistently improves performance on human activity recognition. The features are built using auxiliary atomic action data collected in our lab. We train a kernelized SVM classifier for each atomic action class. Then given a local spatio-temporal cuboid of a test video, we represent it using the responses of our atomic action classifiers. This new atomic action feature is discriminative, and has semantic meanings. We perform extensive experiments on four benchmark action recognition datasets. The results show that atomic action features either outperform the corresponding low level features or significantly boost the recognition performance by combining the two.

1 Introduction

Low level local spatio-temporal features such as HOG and HOF [1–8] have been shown very successful for action recognition in the past. In a "bag of words" representation scheme, these local features are directly clustered to build a visual dictionary and then represented as visual words. During this process, neither semantic nor discriminative cues are utilized. Hence redundant or non-informative visual patterns might be kept. We argue that representing local features in a semantic, discriminative space may offer extra advantages and provide complementary information to that of the low level features.

In this paper, we propose atomic action features, a new representation of local spatio-temporal cuboids based on atomic actions. Atomic actions are basic units of human actions, such as "raising a hand", "one-arm waving". Many atomic actions can be characterized by local motion, and complex actions such as "playing basketball" can be considered as compositions of atomic actions. Intuitively, we can categorize an action based on what atomic actions are observed and how frequent they are. Our idea is to encode local features in atomic action space. Figure 1 illustrate the framework of extracting our atomic action feature representation. The implementation is simple: we train a number of discriminative atomic action classifiers (kernelized SVM classifiers are employed in this paper), then for a local spatio-temporal cuboid, we apply the learned classifiers. A classification score denotes the confidence that a cuboid belongs to an atomic

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 291–300, 2012.

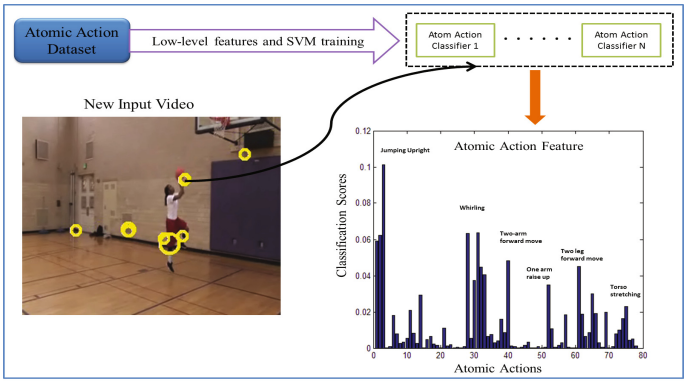

Fig. 1. Illustration of the atomic action feature extraction process. We first collect an atomic action dataset. For each atomic action class, we train a kernelized SVM based on low level features. Given a new action video, for each local spatio-temporal cuboid, we run all the atomic action classifiers. The classification scores are used as the atomic action features to represent the cuboid.

action. Then the set of classification scores are used to represent the cuboid, which shows how likely the cuboid belongs to each of the atomic actions. The new feature representation is complementary to the low-level features, as shown by our experiments.

We build this feature representation based on two insights. First, the representation is discriminative. We train our atomic action SVM classifiers using many positive and negative training examples. In the training process, discriminative visual patterns that are beneficial for classification are preserved, while the others are abandoned. By mapping a local spatio-temporal cuboid to this space, we explicitly exploit its affinity with these discriminative visual patterns. Second, the representation has semantic meanings, and is well aligned with human interpretation.

As a result, though the methodology is conceptually simple, we find it works very well on most of the popular action recognition databases. It either outperforms the corresponding low level features or boosts the performance by combining it with the low level features. And interestingly, our atomic action classifiers have very strong generalization ability. We collect atomic action videos in our lab, with around 10 subjects. The same atomic actions classifiers are applied to four datasets: KTH [9], Hollywood2 [10], Olympic Sport [5] and Youtube [2]. All the results show the effectiveness of our method without adapting the atomic action classifiers. This is very useful, since we don't have to manually annotate atomic action examples for a specific dataset when applying this idea.

1.1 Related Work

Our work is most relevant to the line of work which uses many object categories as the basic representation for image annotation, retrieval, and classification [11, 12, 6, 13]. Our works differs from theirs in two senses. First, we develop this representation for action representation, while they tackle image analysis. Second, our atomic action features are local features. In contrast, [14, 12] build global image features, and [11] detects object instances based on sub-windows, which are still semi-local. In [6], Liu et al. proposed a middle level representation: a video sequence is represented with responses to a set of attribute classifiers. In this paper, we represent each local spatio-temporal cuboid using atomic action features. Local features are expected to be more robust to clutter, occlusion, etc.

We learn atomic action features using positive and negative examples. Recently, there has been growing interest in learning features for action recognition [15, 16]. These works learn spatio-temporal features in a unsupervised manner, to replace the HOG/HOF features, and show promising results. Our work is complementary to theirs, since we can build our atomic action features based on their learned representation.

Atomic actions are studied before by various researchers [17, 18]. They usually aim to reliably detect atomic actions, or build models to model the composition of atomic actions. Different from their work, we encode local features in the semantic atomic action space and can apply it to various action recognition tasks including sports recognition and movie clip recognition.

2 Approach

Our approach is to build atomic action features for human action recognition. The atomic action features are expected to capture the semantic meanings of local spatio-temporal cuboids, and are discriminative. We have training and test videos, we also collect a dataset with atomic action clips. We train a kernelized SVM classifier for each atomic action based on the conventional low level features Then given a local spatio-temporal cuboid, we extract the same low level features, and apply our atomic action classifiers to produce atomic action features, which are the classification responses. We train classifiers based on these new atomic action features. We also combine atomic action features with the original low level features in a multiple kernel learning framework.

2.1 Collecting an Atomic Action Dataset

To our best knowledge, there are no atomic action datasets available. We collect an atomic action dataset in our lab to train the atomic action classifiers. We choose 26 common atomic actions, including “one-hand waving”, “two-hands up”, “stretching”, “stand up”, and so on. For each atomic action class, we invite around 10 volunteers to perform it. Then we can run a saliency detector to detect clean local spatio-temporal cuboid to represent these actions. In order to deal with view variance, we capture each atomic action in three different views. In total, our dataset includes about 1300 videos.

2.2 Training Atomic Action Classifier to Generate Atomic Action Features

We train atomic action classifiers based on local features, as atomic actions are usually characterized by local motion. For each atomic action video, we run the STIP [19] detector to find local spatio-temporal cuboids which contain the salient information. We choose histogram of oriented gradient (HOG) and histogram of optical flow (HOF) [1] to describe local appearance and motion, due to their popularity and the superior performance [1–3, 5]. Following [1], we concatenate HOG and HOF descriptors as a single feature vector For each atomic action category, we randomly choose 2000 cuboids as positive training examples, we also randomly choose 2000 negative samples from all the other categories. A binary SVM classifier with the chi-square kernel is trained based these positive and negative training examples. Note that each atomic action class has three different views. We train a classifier for each view independently due to the big inter-view variation. At the end, we have 78 classifiers in total.

We want to use classification scores as the feature representation, then classification scores of different classifiers must be calibrated. We do this by converting the SVM decision values into probabilistic scores by using the sigmoid mapping function:

$$g(x) = \frac{1}{1 + \exp\left(af(x) + b\right)} \tag{1}$$

where $f(x)$ is the classification score of an atomic action classifier on a local cuboid x, a and b are sigmoid function parameters. We directly use the LIBSVM [20] software to generate probabilistic outputs.

Then given a new local cuboid x_t, we run all the atomic action classifier on it and get an atomic action feature (AAF) vector.

$$AAF(x_t) = [g_1(x_t), \ldots, g_N(x_t)] \tag{2}$$

where $g_i(x)$ is the probabilistic score of the ith atomic action classifier. We use ℓ_2 normalization scheme to normalize the feature vector.

2.3 Using Atomic Action Features to Recognize Actions

We apply our atomic action features to general action recognition. We adopt the most popular "bag of words" scheme to make a fair comparison with the original low level features. But note that our features can also be used with other complex models.

Two types of global representation are tested. The first one is the original bag of words representation. No spatio-temporal information is exploited. In the second method, we follow [1] to partition a video into several spatio-temporal grids. We use three types of spatio-temporal grids: $1 \times 1\ t1$, $1 \times 1\ t2$ and $h3 \times 1\ t1$. More details can be found in [1]. We call the first method BoW and the second methods SPM in the rest of this paper.

Again, we use the SVM classifier with the chi-square kernel to recognize actions.

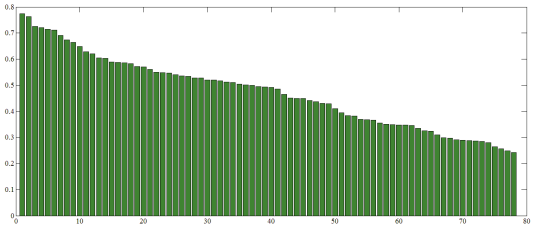

Fig. 2. The average precision (AP) score of the 78 atomic action classifiers on a validation set. The three atomic action classes which have the highest AP values are: torso bending, torso stretching, and one-arm raise up; the three classes which have the lowest AP values are: body whirling, jumping upright and one-arm forward raise up.

For each type of low-level features (such as the HOG/HOF), we can produce the corresponding atomic action features. These two features are expected to be complimentary. We also combine these two features together in the multiple kernel learning framework. We construct a chi-square kernel for each, and then add the two kernels with weights. The weights are learnt via cross validation. Our experimental results show that the combined kernel always works better than the kernel constructed using the original low-level features. This is interesting, as for a low-level feature, we can use this method to boost its performance.

3 Experiments

3.1 Evaluating the Trained Atomic Action Classifiers

We first evaluate the performance of our atomic action classifiers on a validation set. For a particular atomic action classifier, there are 350 positive test samples (cuboids) and 14000 negative test samples (cuboids). An average precision (AP) score is calculated for each atomic action classifier. Figure 2 shows the AP of all categorizes. The mean AP over all classifiers is 0.473. It shows our atomic action classifiers can do reasonably well on classification.

3.2 Performance of the Atomic Action Features on Different Dataset

The same atomic action classifiers are applied to four benchmarks dataset to produce atomic action features for recognition.

The KTH Action Dataset. This dataset is firstly introduced in [20]. We follow a previous experiment setup [9, 3] and train a multi-class classifier and the average accuracy is used to evaluate the performance.

We report the performances of different types of feature on the KTH dataset in Table 1. For all the compared features, the visual dictionary size is 1024.

Table 1. Average Accuracy values on the KTH dataset. "HOG" shows the results of only using the HOG features to represent each local spatio-temporal cuboid. "HOF" shows the results of only using the HOF features to represent each local spatio-temporal cuboid. "HOG/HOF" shows the results of concatenating HOG and HOF features (162 dimensions in total) to represent each local spatio-temporal cuboid. "Atom" shows the results of only using our atomic action features to represent each local spatio-temporal cuboid. "HOG/HOF+Atom" shows the results of combing "HOG/HOF" and atomic action features with a multiple kernel SVM classifier. "Bow" means the standard bag of word scheme; "SPM" means spatio-temporal grids are used, and each grid is represented as a bag of words histogram.

Feature	HOG	HOF	HOG/HOF	Atom	HOG/HOF+Atom
BoW	81.1%	91.4%	89.6%	88.3%	93.2%
SPM	81.6%	90.7%	88.5%	87.5%	93.0%

We use the same size for all the other four datasets. And a chi-square kernelized SVM classifier is applied. The performance of atomic action features is comparable to that of "HOF/HOF". Combining the two significantly boosts the performance.

The Hollywood2 Dataset. This is a dataset of 12 action classes collected from 69 Hollywood movies [10]. We follow the experiment setup of [3] and train a binary classifier for each action class. We first compute the average precision (AP) for each action class. And the mean average precision over all the action classes is reported, as in [10, 3].

Table 2. Average Precision (AP) values on the Hollywood2 Dataset. (Please refer to table 1 for notation definition.)

Feature	HOG	HOF	HOG/HOF	Atom	HOG/HOF+Atom
BoW	31.8%	40.3%	41.3%	43.1%	46.3%
SPM	37.6%	42.2%	44.0%	45.9%	49.4%

A comparison of our atomic action feature against other feature for each action category on the HOllywood2 dataset is shown in Table 2. For both BOW and SPM methods, our atomic action feature outperform HOG, HOF, and the corresponding HOG/HOF features. Atomic action features outperform the others on 7 categories with BOW, and on 8 categories with SPM over all the action classes. Combining "HOG/HOF" and atomic actions features obtains around 5% improvement on mean average precision, for both the BOW and SPM methods.

Olympic Sports Dataset. This dataset is created by Niebles et al. [5]. We follow their experimental setting in [5] and train a binary classifier for each action class. Similar to the Hollywood2 dataset, average precision (AP) is calculated for each action class, and mean average precision values over all the action classes are reported.

Table 3. Average Precision (AP) values on the Olympic Sports Dataset. (Please refer to table 1 for notation definition.

Feature	HOG	HOF	HOG/HOF	Atom	HOG/HOF+Atom
BoW	55.1%	57.2%	59.2%	63.1%	68.4%
SPM	61.9%	59.5%	63.9%	64.5%	71.0%

Table 4. Average Accuracy values for the classification task in YouTube Action dataset.(Please refer to table 1 for notation definition.)

Feature	HOG	HOF	HOG/HOF	Atom	HOG/HOF+Atom
BoW	61.7%	56.0%	61.9%	59.3%	68.4%
SPM	65.9%	57.7%	65.9%	62.7%	72.7%

In table 3, we show the performance of different features on the Olympic Sport dataset. Our atomic action feature achieve the best results on the mean average precision over all the categories, compared to HOG, HOF, and HOG/HOF. This shows our atomic action representation is very discriminative on this dataset. On most categorizes (12/16 for BOW and 11/16 for SPM), our atomic action features outperform the corresponding low-level HOG/HOF features. Combing the two types of features results in about 9% and 7% gain in mean AP for BOW and SPM methods, respectively. We shows two examples of our atomic action features on the Olympic Sport dataset in Figure 3, which are very descriptive.

YouTube Action Dataset. This dataset is published in [2] for evaluating action recognition in unconstrained videos. We follow their leave on out cross validation (LOOCV) method for these 25 groups in our experiments. Average accuracy scores over all the classes are compared.

The results are compared in table 4. From the table, we can see a gain of about 7% is achieved by combining our atomic action features with "HOG/HOF", compared to only using "HOF/HOF". Interestingly, "HOF/HOF+Atom" outperforms a more complicated approach proposed by Liu et al. [2], whose average accuracy number is 71.2%.

Table 5. Comparison of using the "HOG/HOF + Atom" feature with other methods in the literature

KTH		Olympic Sports		Hollywood2	
Niebles et al. [5]	91.3%	Niebles et al. [5]	72.1%	Alexander et al. [21]	45.3%
Laptev et al. [1]	91.8%	Liu et al. [6]	74.3%	Laptev et al.[1]	47.7%
Liu et al. [6]	91.6%				
Our Method	93.2%	Our Method	71.0%	Our Method	49.4%

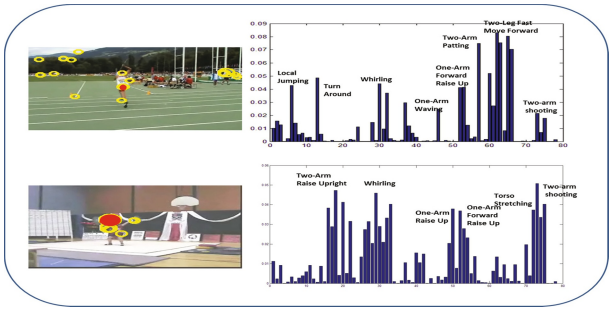

Fig. 3. Two examples of our atomic action feature on the Olympic Sports dataset. Left : local spatio-temporal cuboid (indicated in red), right : the corresponding atomic action features.

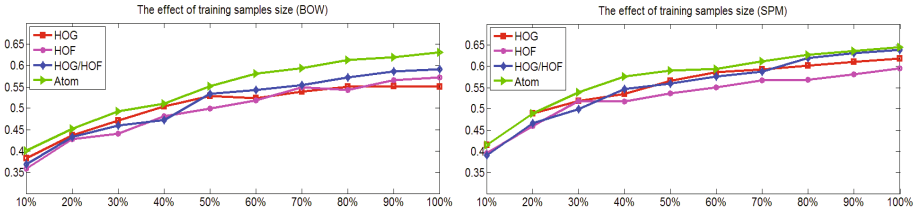

Fig. 4. The performance of different features on the Olympic Sports dataset, with different number of training samples. Our atomic action feature representation always outperform HOG, HOF and HOG/HOF features.

3.3 Comparison with Previous Work

In table 5, we compare our results with those of several previous papers on the KTH, Olympic Sports, and Hollywood2 datasets. Even only using a less sophisticated model (SPM), we find our approach "HOG/HOF+atom" works reasonably well compared to many previous, more sophisticated models.

3.4 The Effect of Training Sample Size

In this section, we investigate the effect of training sample size, when using the proposed atomic action features. We test the performance of atomic action features with different number of training samples on the Olympic Sports dataset. We randomly select 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of the positive and negative videos respectively in the training data for each category to do the experiments. For each size, we repeat the experiments 10 times by randomly selecting training examples. The results are averaged and compared in Figure 4. We can see from the figure that the atomic action features always perform better than the other features (HOG, HOF, and HOG/HOF), with varying number of training samples.

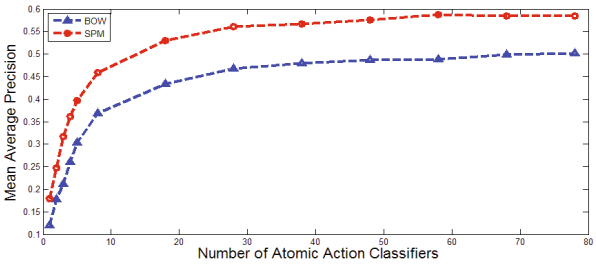

Fig. 5. The performance with different atomic action classifiers sizes on Olympic Sports dataset. For each classifier size, we repeat the experiments 30 times by randomly selecting classifiers. Averaged score is reported and compared.

3.5 The Effect of Atomic Action Classifier Size

We also investigate the effect of the number of atomic action classifiers, on the Olympic Sports datasets. We perform experiments with different numbers of atomic action classifiers: 1, 2, 3, 4, 5, 8, 18, 28, 38, 48, 58, and 68 respectively. For each number, we repeat the experiments 30 times, by randomly choosing a subset of atomic action classifiers. Averaged results are reported. Figure 5 shows the mean average precision values with varying number of atomic action classifiers. The improvement is not so significant when the size of atomic action classifiers reaches 30.

4 Conclusions and Discussions

In this paper, we have presented a simple method to build atomic action features for action recognition. Our extensive results on four action recognition benchmark datasets show the effectiveness of this method. There are two interesting things about this new type of feature. First, for a state-of-the-art low level feature (HOG/HOF), our method can at least help improving its performance by combing it with the atomic action features. Second, our atomic action classifiers have very strong generalization ability because they only capture local motion information. We build the atomic action classifiers using a dataset collected in our lab, but can successfully apply it to various datasets.

Acknowledgments. This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

References

1. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. CVPR (2008)

2. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: Proc. CVPR (2009)
3. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: Proc. BMVC (2009)
4. Ni, B., Yan, S., Kassim, A.A.: Recognizing human group activities with localized causalities. In: Proc. CVPR (2009)
5. Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)
6. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Proc. CVPR (2011)
7. Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: ICCV Workshops (2011)
8. Zhang, T., Xu, C., Zhu, G., Liu, S., Lu, H.: A generic framework for event detection in various video domains. In: ACM Multimedia (2010)
9. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proc. ICPR (2004)
10. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: Proc. CVPR (2009)
11. Li, L.-J., Su, H., Lim, Y., Fei-Fei, L.: Objects as attributes for scene classificcation. In: ECCV Workshop (2010)
12. Rasiwasia, N., Vasconcelos, N.: Scene classification with low-dimensional semantic spaces and weak supervision. In: Proc. CVPR (2008)
13. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: Proc. CVPR (2012)
14. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient Object Category Recognition Using Classemes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 776–789. Springer, Heidelberg (2010)
15. Taylor, G.W., Fergus, R., LeCun, Y., Bregler, C.: Convolutional Learning of Spatio-temporal Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 140–153. Springer, Heidelberg (2010)
16. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Proc. CVPR (2011)
17. Gaidon, A., Harchaoui, Z., Schmid, C.: Actom sequence models for efficient action detection. In: Proc. CVPR (2011)
18. Ryoo, M.S., Aggarwal, J.K.: Recognition of composite human activities through context-free grammar based representation. In: Proc. CVPR (2006)
19. Laptev, I.: On space-time interest points. IJCV 64(2-3), 107–123 (2005)
20. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
21. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: Proc. BMVC (2008)