Appendix
Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis

Quoc V. Le, Will Y. Zou, Serena Y. Yeung, Andrew Y. Ng

quocle@cs.stanford.edu, wzou@cs.stanford.edu, syyeung@stanford.edu, ang@cs.stanford.edu

October 3, 2011

1 Introduction

In the experiment section of the paper, we present the performance of the convolutional ISA model on 4 human action recognition datasets. Due to the space limitations in our paper, we clarify the setup, parameters, and evaluation metric of experiments in this Appendix.

For all the datasets, we use the same pipeline with Wang et. al [5]. This pipleline first extracts features (or in the case Wang et. al [5], descriptors such as HOG3D) from videos on a dense grid in which cube samples overlap 50% in x, y and t dimensions. K-means vector quantization is applied on the extracted features and each video is histogramed to form a bag-of-words representation. Finally, the bag-of-words representation is L1-normalized and a χ^2 kernel SVM is used to classify human action. To make our experiments comparable to earlier work, we apply the same evaluation setting and metric as prior art in each dataset. We describe these metrics in the following sections. Detailed results, such as average precision/accuracy per action class and confusion matrices, are also provided.

2 Results with respect to number of K-means examples

The results presented in the paper use the same pipleline as Wang et. al [5]. We clarify in this section the difference in results between when we use the same number of K-means examples as in [5] (100,000), and when we allow a large number of K-means examples (3,000,000). This comparison is given in Table 1. The best results are given in bold.

Table 1: Results with respect to number of K-means samples (cells with '-' indicates that results were not reported in the corresponding literature)

	Hollywood2	KTH	UCF	Youtube
Wang et. al [5]	47.7%	92.1%	85.6%	-
Liu et. al [2]	-	-	-	71.2%
100,000 K-means samples	50.8%	**93.8%**	86.5%	75.6%
3,000,000 K-means samples	**53.3%**	-	**86.8%**	**76.5%**

The results increase as we allow larger number of K-means examples. As can be seen from the above comparison, the results using our method and exactly the same settings as Wang et. al [5] out-performs the prior state-of-the-art results. This illustrates the of the features from our unsupervised learning method. In the following sections we provide details of the best results shown in bold in Table 1.

3 Experiment settings and detailed results

In the following subsections we report experiment settings, detailed classification results and confusion matrices on 4 action recognition datasets: Hollywood2, KTH, UCF and Youtube. To the best of our knowledge, we compare results from our method to per-action-class and average classification results previously reported. In the summary tables, cells with a '-' indicates that results were not reported in the corresponding literature.

3.1 Hollywood2

The Hollywood2 human actions dataset (http://pascal.inrialpes.fr/hollywood2/) containing 823 train and 872 test video clips. There are in total 12 action classes and each video clip may have more than one action label. We train 12 binary SVM classifiers, one for each action. For evaluation, the final average precision(AP) metric is obtained by taking the average of AP for each classifier run on the test set.
In Figure 2 we compare the result using our method to prior art on the Hollywood2 dataset.

Table 2: Hollywood2: average precision by action

	Marszalek et. al [3]	Wang et. al [5]	Our method
Answer phone	13.1%	-	**29.9%**
Drive car	81%	-	**85.2%**
Eat	30.6%	-	**59.7%**
Fight person	62.5%	-	**77.2%**
Get out of car	8.6%	-	**45.4%**
Hand shake	19.1%	-	**20.3%**
Hug person	17.0%	-	**38.2%**
Kiss	**57.9%**	-	**57.9%**
Run	55.5%	-	**75.7%**
Sit down	30%	-	**59.4%**
Sit up	17.8%	-	**25.7%**
Stand up	33.5%	-	**64.7%**
Average	35.5%	47.7%	**53.3%**

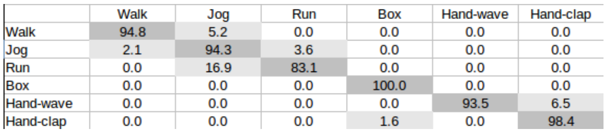
3.2 KTH

KTH actions dataset(http://www.nada.kth.se/cvap/actions/) contains 2391 video samples with 6 action labels. We follow [4] and [5] and split the dataset into the test set, which contains subjects 2, 3, 5, 6, 7, 8, 9, 10, 12) and the training set, which contains the rest of the subjects. For evaluation, we train a multi-class SVM and evaluate on the test set. Detailed results with comparison to prior art is given in Table 3. The confusion matrix is provided in Figure 1.

Table 3: KTH: average accuracy by action class

	Schuldt et. al [4]	Kläser et. al [1]	Wang et. al [5]	Our method
Walk	83.8%	-	-	**94.8%**
Jog	60.4%	-	-	**91.2%**
Run	54.9%	-	-	**85%**
Box	97.9%	-	-	**100%**
Hand wave	73.6%	-	-	**94.6%**
Hand clap	59.7%	-	-	**97.3%**
Average	71.7%	91.4%	92.1%	**93.8%**

If we use 75% overlap grid sampling as compared to 50% in [5], our method achieves a classification accuracy of 94.5% on the KTH dataset.



	Walk	Jog	Run	Box	Hand-wave	Hand-clap
Walk	94.8	5.2	0.0	0.0	0.0	0.0
Jog	2.1	94.3	3.6	0.0	0.0	0.0
Run	0.0	16.9	83.1	0.0	0.0	0.0
Box	0.0	0.0	0.0	100.0	0.0	0.0
Hand-wave	0.0	0.0	0.0	0.0	93.5	6.5
Hand-clap	0.0	0.0	0.0	1.6	0.0	98.4

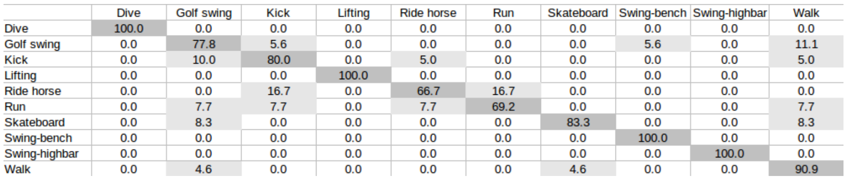
Figure 1: Confusion matrix for the KTH dataset

3.3 UCF

UCF sport actions dataset(http://server.cs.ucf.edu/∼vision/data.html) contains 150 video samples with 10 action labels. As in [5], we extend the dataset by adding a horizontally flipped version of each video. For evaluation, we perform classification with a multi-class SVM using leave-one-out. This means for each clip in the dataset, we predict its label while training on all other clips, except for the flipped version of the tested video clip. The detailed results on UCF is given in Table 4. The confusion matrix is provided in Figure 2.

Table 4: UCF: average accuracy by action class

	Wang et. al [5]	Our method
Dive	-	100%
Golf swing	-	77.8%
Kick	-	80%
Lifting	-	100%
Ride horse	-	66.7%
Run	-	69.2%
Skateboard	-	83.3%
Swing-bench	-	100%
Swing-highbar	-	100%
Walk	-	90.9%
Average	85.6%	**86.8%**



	Dive	Golf swing	Kick	Lifting	Ride horse	Run	Skateboard	Swing-bench	Swing-highbar	Walk
Dive	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Golf swing	0.0	77.8	5.6	0.0	0.0	0.0	0.0	5.6	0.0	11.1
Kick	0.0	10.0	80.0	0.0	5.0	0.0	0.0	0.0	0.0	5.0
Lifting	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Ride horse	0.0	0.0	16.7	0.0	66.7	16.7	0.0	0.0	0.0	0.0
Run	0.0	7.7	7.7	0.0	7.7	69.2	0.0	0.0	0.0	7.7
Skateboard	0.0	8.3	0.0	0.0	0.0	0.0	83.3	0.0	0.0	8.3
Swing-bench	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Swing-highbar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
Walk	0.0	4.6	0.0	0.0	0.0	0.0	4.6	0.0	0.0	90.9

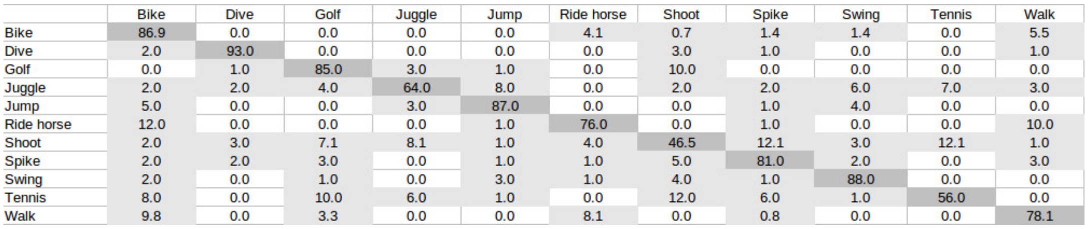
Figure 2: Confusion matrix for the UCF dataset

3.4 Youtube

The more recent Youtube actions dataset(http://server.cs.ucf.edu/∼vision/data.html) contains 1600 video clips with 11 actions. We use the experiment setting in [2], which takes part of the dataset(1168 videos, including all videos from biking and walking classes, and only videos from indexed 01 to 04 for the rest of classes) and obtain the average accuracy over 25-fold cross-validation. The 25-fold cross-validation is performed according to the authors' original split. Table 5 and Figure 3 provides per class accuracy and the confusion matrix, respectively.

Table 5: Youtube: average accuracy by action class

	Liu et. al	Our method
Cycle	73%	**86.9%**
Dive	81%	**93%**
Golf	**86%**	85%
Juggle	54%	**64%**
Jump	79%	**87%**
Ride horse	72%	**76%**
Basketball shoot	**53%**	46.5%
Volleyball spike	73.3%	**81%**
Swing	57%	**88%**
Tennis	**80%**	56%
Walk	75%	**78.1%**
Average	71.2%	**76.5%**



	Bike	Dive	Golf	Juggle	Jump	Ride horse	Shoot	Spike	Swing	Tennis	Walk
Bike	86.9	0.0	0.0	0.0	0.0	4.1	0.7	1.4	1.4	0.0	5.5
Dive	2.0	93.0	0.0	0.0	0.0	0.0	3.0	1.0	0.0	0.0	1.0
Golf	0.0	1.0	85.0	3.0	1.0	0.0	10.0	0.0	0.0	0.0	0.0
Juggle	2.0	2.0	4.0	64.0	8.0	0.0	2.0	2.0	6.0	7.0	3.0
Jump	5.0	0.0	0.0	3.0	87.0	0.0	0.0	1.0	4.0	0.0	0.0
Ride horse	12.0	0.0	0.0	0.0	1.0	76.0	0.0	1.0	0.0	0.0	10.0
Shoot	2.0	3.0	7.1	8.1	1.0	4.0	46.5	12.1	3.0	12.1	1.0
Spike	2.0	2.0	3.0	0.0	1.0	1.0	5.0	81.0	2.0	0.0	3.0
Swing	2.0	0.0	1.0	0.0	3.0	1.0	4.0	1.0	88.0	0.0	0.0
Tennis	8.0	0.0	10.0	6.0	1.0	0.0	12.0	6.0	1.0	56.0	0.0
Walk	9.8	0.0	3.3	0.0	0.0	8.1	0.0	0.8	0.0	0.0	78.1

Figure 3: Confusion matrix for the Youtube dataset

4 Typos in the main paper

In the paper, there is a typo in the definition of the activation of the hidden units. In particular, in section 3.1 we defined the activation of the hidden unit to be

$$p_i(x^t; W, V) = \sqrt{\sum_{k=1}^{m} V_{ik} \left(\sum_{j=1}^{n} W_{kj} x_j^t \right)^2}. \tag{1}$$

The correct definition of the activation of the hidden units should be

$$p_i(x^t; W, V) = \sqrt{\sum_{l=1}^{k} V_{il} \left(\sum_{j=1}^{n} W_{lj} x_j^t \right)^2}. \tag{2}$$

Here, as stated in the paper, $V \in \mathbb{R}^{m \times k}$ and $W \in \mathbb{R}^{k \times n}$.

References

[1] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D gradients. In *BMVC*, 2008.

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

[3] M. Marzalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

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

[5] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2010.