Exploring Dense Trajectory Feature and Encoding Methods for Human Interaction Recognition

Xiaojiang Peng[1,2]
[1]Southwest Jiaotong University, Chengdu, China, 610031
xiaojiangp@gmail.com

Qiang Peng
Southwest Jiaotong University
Chengdu, China, 610031
qpeng@home.swjtu.edu.cn

Yu Qiao*
[2]Shenzhen Key Lab of CVPR, Shenzhen Institute of Advanced Technology, CAS
yu.qiao@siat.ac.cn

Xiao Wu
Southwest Jiaotong University
Chengdu, China, 610031
wuxiaohk@home.swjtu.edu.cn

Xianbiao Qi
Beijing University of Posts and Telecommunications
Beijing, China
qixianbiao@gmail.com

Yanhua Liu
Samsung Guangzhou Mobile R&D Center
Guangzhou, China, 510663
yanhua.liu@samsung.com

ABSTRACT

Recently, human activity recognition has obtained increasing attention due to its wide range of potential applications. Much progress has been made to improve the performance on single actions in videos while few on collective and interactive activities. Human interaction is a more challenging task owing to multi-actors in an execution. In this paper, we utilize multi-scale dense trajectories and explore four advanced feature encoding methods on the human interaction dataset with a bag-of-features framework. Particularly, dense trajectories are described by shape, histogram of gradient orientation, histogram of flow orientation and motion boundary histogram, and all these are computed by integral images. Experimental results on the UT-Interaction dataset show that our approach outperforms state-of-the-art methods by 7-14%. Additionally, we thoroughly analyse a finding that the performance of vector quantization is on par with or even better than other sophisticated feature encoding methods by using dense trajectories in videos.

Categories and Subject Descriptors

I.4 [**Image Processing and Compute Vision**]: Applications; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*human activity recognition*

General Terms

Algorithms, Experimentation, Theory

*Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICIMCS'13, Aug. 17–19, 2013, Huangshan, Anhui, China.
Copyright 2013 ACM 978-1-4503-2252-2/13/08 ...$15.00.

Keywords

Dense trajectory, bag-of-features, feature encoding, human activity recognition

1. INTRODUCTION

Automatic recognition of human activity in videos has been an active research area in recent years due to its wide potential applications, such as smart video surveillance, video indexing and human-computer interface. Though various approaches have been proposed and many progresses have been achieved, it still remains a challenging task due to its large intra-variations, clutter and occlusion in background or foreground and other fundamental difficulties.

With the development of human activity recognition, many activity datasets are introduced. Weizmann dataset [1] and KTH dataset [15] are basic benchmarks for kinematics activities whose backgrounds are homogenous. Some datasets like Hollywood2 [12], UCF Youtube[1], and HMDB51 [7] have been used for evaluation in realistic environments. Among these videos, most of them have only one actor. Another kind of benchmark is group activities, such as Collective[2] and UT-Interaction[3] [14] datasets. UT-Interaction dataset is a pairwise human activity dataset, we will give the details in Section 3.

Among state-of-the-art methods, local space-time feature with bag-of-features (BoF) framework is a successful representation for action recognition. Laptev [8] has introduced space-time interest points (STIP) by extending the Harris detector to video. Dollar et al. [5] detected space-time salient points by 2D spatial Gaussian and 1D temporal Gabor filters. Feature descriptors range from higher order derivatives, gradient information, optical flow, and brightness information [5] to spatio-temporal extensions of image descriptors, such as 3D-SIFT [16], HOG3D [6], extended SURF [21], and Local Trinary Patterns [23]. Lately, Wang [18, 19] has demonstrated the powerful ability of dense trajectory (DT) with vector quantization (VQ) on most of the

[1]http://crcv.ucf.edu/data/UCF_YouTube_Action.php
[2]http://www.eecs.umich.edu/vision/activity-dataset.html
[3]http://cvrc.ece.utexas.edu/SDHA2010/index.html

single human activity datasets. However, due to the huge amount of trajectories and expensive computation, the performance of those well-designed feature encoding methods like sparse coding (SC) [10], locality-constrained linear coding (LLC) [20] and locality soft-assignment (LSA) [11] are unknown or undiscussed. In this paper, we thoroughly explore dense trajectory and these encoding methods with the BoF framework on UT-Interaction.

The main contributions of this paper can be summarized into two folds. First, we utilize DT with four advanced feature encoding methods for the human interaction dataset and give a comprehensive comparison. Second, we thoroughly analyse the finding that local encoding methods and max-pooling approach will degrade by using a finite visual codebook in the face of so many dense trajectories.

Section 2 briefly introduces our framework and related methods. Section 3 shows the experimental results and gives a comprehensive discussion. We conclude our paper in Section 4.

2. FRAMEWORK AND APPROACHES

An illustration of our framework is shown in Figure 1. First, multi-scale dense trajectories are extracted from raw videos and four descriptors are computed and combined for each trajectory. Descriptors are DT shape, histogram of gradient orientation (HoG), histogram of flow orientation (HoF) and motion boundary histogram (MBH). Then, a visual codebook is learned from the training samples using k-means. After that, every combined descriptor is encoded and a coding coefficient vector is yielded. We explore four methods in this step: vector quantization, sparse coding , locality-constrained linear coding and localized soft-assignment. Later, coding coefficient vectors in a single video are pooled into a global histogram. Finally, a one-vs-rest χ^2 kernel SVM classifier is trained and applied to predict testing videos.

In the following, we describe dense trajectory method in details and review feature encoding and pooling methods we use.

2.1 Dense Trajectory

Inspired by dense sampling in image classification, a simply-designed dense trajectory method is introduced [18]. It contains the following steps.

Dense sampling and filtering. Feature points are sampled in the current frame on a grid by a step size w at S spatial scales. To track successfully, points in homogeneous image areas are filtered out by checking the eigenvalues of their auto-correlation matrix.

Trajectories. Dense points are tracked by median-filtered optical flow on each spatial scale separately. Tracked points in successive frames at scale s are concatenated to form trajectories: $(P_t^s, P_{t+1}^s, ...)$, where $P_t^s = (x_t^s, y_t^s)$ is the spatial position. Some trajectories are shown in Fig. 1. To prevent the trajectories from drifting, the length is limited to L frames. Once a trajectory's length reaches L, its mean position drift and variation will be checked. Trajectories with tiny or large mean drift and variation will be pruned since those are static or erroneous trajectories.

Descriptors for DT. There are four types of descriptors for each trajectory [19]. The trajectory shape is described by a sequence $(\Delta P_t^s, ..., \Delta P_{t+L}^s)$ of displacement vectors $\Delta P_t^s = (x_{t+1}^s - x_t^s, y_{t+1}^s - y_t)$. Usually, this vector is normalized by the ℓ_1-norm. Therefore, we obtain a $2L$ dimensional shape descriptor. To catch the motion and structure information, HoG, HoF and MBH are extracted within a space-time volume whose size is $N \times N \times L$ aligned with the trajectory. HoG and HoF descriptors are popular methods which yielded excellent results on many datasets [9]. The MBH represents the gradient of the optical flow which is initially introduced in human detection [4]. To embed more structure information, we usually subdivide the volume into a spatio-temporal grid of size $n_\sigma \times n_\sigma \times n_\tau$ as shown in Fig. 1. Assume n bins are used for HoG and MBH, and $n+1$ (1 for static) bins for HoF, then we get a $n_\sigma \times n_\sigma \times n_\tau \times n$ vector for HoG, $n_\sigma \times n_\sigma \times n_\tau \times (n+1)$ for HoF and $n_\sigma \times n_\sigma \times n_\tau \times n \times 2$ for MBH (2 corresponds to the horizontal and vertical flow components).

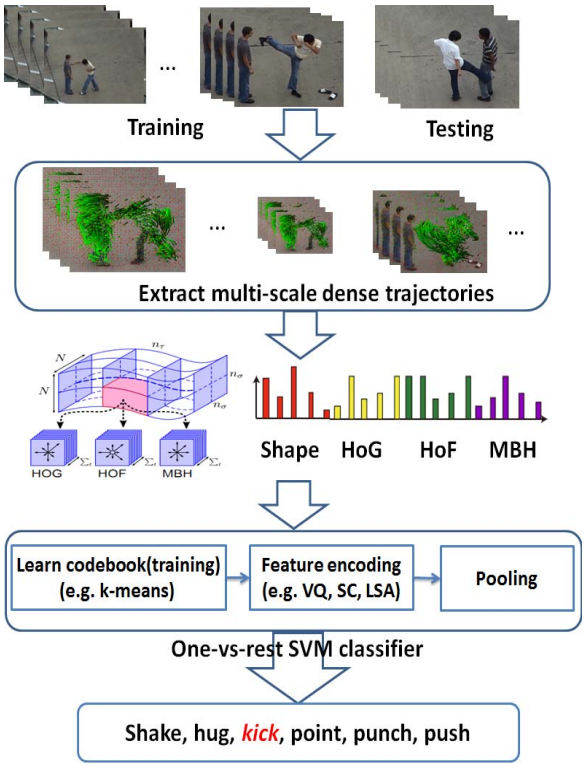

Figure 1: A schematic framework of our work.

2.2 Feature Encoding and Pooling

Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \mathcal{R}^{d \times n}$ be a set of feature descriptors. Given a codebook $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k] \in \mathcal{R}^{d \times k}$. The purpose of encoding is to compute a coefficient for input $\mathbf{y}$ with $\mathbf{D}$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathcal{R}^{k \times n}$ be the coefficient vectors. Following are different encoding methods.

Vector quantization. VQ is the standard encoding method of BoF, which solves the following constrained objective function:

$$arg \min_{\mathbf{x}} ||\mathbf{y} - \mathbf{D}\mathbf{x}||^2 \qquad s.t. ||\mathbf{x}||_{\ell_0} = 1 \tag{1}$$

where the constraint $||\mathbf{x}||_{\ell_0} = 1$ means that there will be only one non-zero element in $\mathbf{x}$.

Localized soft-assignment. Here we explore the k-nearest neighborhood or "localized" soft-assignment proposed

Figure 2: Example frames of six pairwise human activities.

in [11]. Let x_i be an element of vector $\mathbf{x}$,

$$x_i = \begin{cases} \frac{\exp(-\beta||\mathbf{y}-\mathbf{d}_i||_2^2)}{\sum_{i=1}^{K}\exp(-\beta||\mathbf{y}-\mathbf{d}_i||_2^2)} & \text{if } \mathbf{d}_i \in N_k(\mathbf{y}) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $N_k(\mathbf{y})$ denotes the k-nearest neighborhood of $\mathbf{y}$. β is a smoothing factor controlling the softness of the assignment.

Sparse coding. SC tries to minimize the reconstruction error with the least number of words in $\mathbf{D}$. The general form of sparse coding is as following equation,

$$\mathbf{x} = arg\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{D}\mathbf{x}||^2 + \lambda\psi(\mathbf{x}) \quad (3)$$

where $\psi(x)$ is the sparse constraint and λ is a sparse factor. There are two formats for the constraint, the ℓ_0-norm and ℓ_1-norm. Here we apply ℓ_1-norm by using "feature-sign" method in [10].

Locality-constrained linear coding (LLC). LLC [20] suggested that locality is more essential than sparsity. The coefficient vector is obtained by solving the following optimization:

$$\mathbf{x} = arg\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{D}\mathbf{x}||^2 + \lambda||\mathbf{e} \odot \mathbf{x}||^2, \quad s.t. \quad \mathbf{1}^\top\mathbf{x} = 1 \quad (4)$$

where $\mathbf{e} = \exp(\text{dist}(\mathbf{y}, \mathbf{D})/\sigma)$ and $\text{dist}(\mathbf{y}, \mathbf{D})$ denotes the Euclidean distance vector between $\mathbf{y}$ and the words of $\mathbf{D}$. σ is a parameter controlling the weight vector $\mathbf{e}$. In our experiments, we apply the k-NN version of LLC, which is an approximation with low computational complexity in practice.

Pooling. To obtain a global representation for each video, all the encoding coefficient vectors of feature descriptors in individual video are pooled to yield a global histogram. Those vectors can be pooled in one of two ways: sum pooling, in which case they are combined additively, or max pooling, in which case each entry of the global histogram is assigned a value equal to the maximum across all feature encoding coefficient vectors. We use max pooling for the LLC encoding, sum pooling for VQ, and both for the others.

3. EXPERIMENTS

We use the UT-Interaction dataset for our exploration which is a standard human interaction dataset. It contains 6 classes of pairwise human activities: hand-shake, point, hug, push, kick and punch. Fig. 2 shows example frames of these pairwise-person activities. There are a total of 120 clips. A number of actors with more than 15 different clothing conditions appear in the videos. The clips are divided into two sets. The set #1 is composed of 60 clips taken on a parking lot. The set #2 (i.e. the other clips) are taken at a lawn on a windy day. More camera jitters occur in set #2. We follow the setting of [14] that uses 10-fold leave-one-out cross validation per set, and report the average class accuracy. In particular, we set the parameters $(w, S, N, L, n_\sigma, n_\tau, n)$ mentioned in Section 2 to be $(8, 8, 32, 15, 2, 3, 8)$, so the dimension of combined vector is 426. Codebooks are generated 10 times using 100k features sampled from training set randomly and uniformly.

Table 1: Results of different methods and codebook size on UT-Interaction set #1

Size	100	500	1k	2k	4k	6k
VQ_s	85.50	90.50	**91.33**	90.50	**92.17**	90.50
LLC_m	73.50	87.12	89.67	85.55	85.55	85.55
LSA_s	83.83	87.17	**91.33**	89.67	**91.33**	**91.33**
LSA_m	51.17	59.83	68.50	84.67	85.55	85.55
SC_s	82.17	86.33	**91.33**	**91.33**	**92.50**	**92.50**
SC_m	60.00	86.33	90.50	87.17	90.50	90.50

Table 2: Results of different methods and codebook size on UT-Interaction set #2

Size	100	500	1k	2k	4k	6k
VQ_s	81.67	85.00	**91.67**	90.00	88.33	**91.67**
LLC_m	71.67	81.67	**88.33**	**88.33**	**88.33**	**88.33**
LSA_s	73.33	81.67	81.67	**88.33**	**88.33**	**91.67**
LSA_m	41.67	58.33	75.00	78.33	85.00	85.00
SC_s	76.67	86.67	**88.33**	**91.67**	**90.00**	**91.67**
SC_m	73.33	81.67	83.33	86.67	82.17	90.00

3.1 Experimental Results and Analysis

The first purpose of our experiments is to explore the codebook size and pooling strategies with different feature encoding methods. And the combined descriptor is applied. We use the default parameters for all encoding approaches that is 5 -NN for LLC, L-SA and SC (a approximate version of SC provided in [22]). Both max pooling and sum or average pooling are used for L-SA and SC. The results on set #1 and #2 are reported in Table 1 and 2. All numbers are percent accuracy. "_s" and "_m" represent sum-pooling and max-pooling, respectively. The top-3 performances are bold.

From these results, a handful of trends are readily apparent. First, we note that sum-pooling is always superior to max-pooling on both sets. To validate this result, we also perform the similar experiments on KTH dataset and the same trend is given. We will not exhibit here due to limited space. However, this is not the case in image classification using dense features [2]. Considering image classification, images are usually forced to a size of 512 pixels at the max side, and dense features are extracted on a grid by a step size 8 in which case about 4k features are generated. But in our experiments, the max number of dense trajectories in a single clip can reach 40k and the intrinsic foreground variations between training and testing samples are very large because we do not generate the codebook on whole data but training features only. As the theoretical analysis in [2],

Table 3: Results of different k for L-SA with the codeboke size 1k on set #2 of UT-Interaction

k	20	10	5	4	2	1
LSA_s	78.33	78.33	81.67	83.33	86.67	91.67

Table 4: Results of different descriptors on both sets of UT-Interaction

	Shape	HoG	HoF	MBH	All-combined
VQ_s#1	76.33	83.33	83.83	**93.33**	91.33
SC_s#1	73.83	77.67	82.17	**94.50**	91.33
VQ_s#2	70.00	63.33	86.67	86.67	**91.67**
SC_s#2	71.67	65.00	80.00	85.00	**88.33**

when the pool cardinality is large, average or sum pooling is robust to intrinsic foreground variability, while it is not the case with max pooling. Second, the best sizes of codebook for both set may be 4k and 6k from the results.

Another striking result is that VQ is always on par with or even better than other sophisticated encoding methods even though sum pooling is used. In the following we will discuss the essence that why this happens.

3.1.1 *Analysis in Data Manifold*

A fundamental assumption of "localized" encoding methods like LLC and L-SA is that local features reside on a lower-dimensional manifold in an ambient descriptor space. The presence of a manifold structure suggests that the Euclidean distance we used in k-NN is only meaningful within a local region. Out of this region, two local features measured close by the Euclidean distance might be actually far from each [11]. That is to say we should not use those words d which are out of the region to encode a new feature. But how local is the desired region used for encoding a testing feature? In other words, what the idea k is.

As shown in Table 2, VQ is significantly better than others when the size of codebook is 1k. We plot the words of this codebook and some testing features randomly sampling from different classes onto the 2D space via multidimensional scaling in Fig. 3. From Fig. 3, we can see that most of the testing features are not close with a number of words, not 5 at least. So as mentioned above, we are likely to get worse performance if we use the default 5-NN in LLC, L-SA and SC. When the size of codebook increases, the manifold structure becomes more dense, and the testing features are likely to close with a number of words. That is why the results get better when the codebook size increases in Table 1 and Table 2.

3.1.2 *Make the Encoding More "localized"*

To demonstrate the above conclusion, we test different k using L-SA for an instance. The results are reported in Table 3. The trend is clear that the results become better with the descent of k. It is the same with VQ when k decreases to 1.

3.2 Descriptors

As in [19], we also evaluate different descriptors of DT on both sets of UT-Interaction dataset. To save computation, we only explore VQ and SC_s with the size of codebook 1k since they perform well at most of the time.

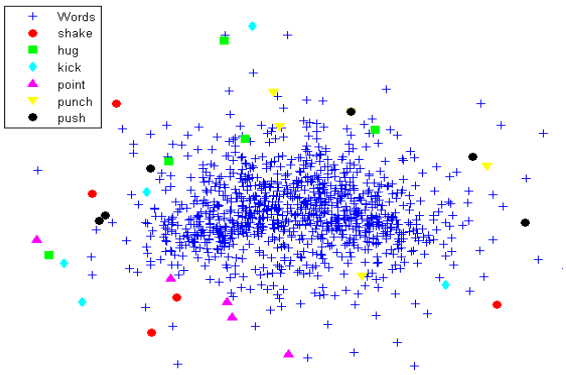

Figure 3: All words of the codebook with size 1k in Table 2 and testing features sampled from each category in the 2D space via multidimensional scaling.

Table 5: Performance comparison on UT-Interaction. "-" represents no reported results.

	shake	hug	kick	point	punch	push	total
[14] #1	0.8	0.9	0.9	1	0.56	0.73	0.85
[17] #1	0.7	1	1	1	0.7	0.9	0.88
[13] #1	-	-	-	-	-	-	0.85
Ours #1	1	1	1	1	0.67	1	**0.945**
[14] #2	0.8	0.8	0.6	0.9	0.7	0.4	0.7
[17] #2	0.5	0.9	1	1	0.8	0.4	0.77
[3] #2	-	-	-	-	-	-	0.78
Ours #2	0.9	1	1	1	0.7	0.9	**0.917**

Results are shown in Table 4. The above two rows are results on set #1 and the others on set #2. For single descriptor, motion boundary histogram is more discriminative than others due to its robustness to camera motion especially on set #1. For the set #1, only clearly motions occur in the foreground. For the set #2, there are some clutter motion in the background due to the windy day, so combined descriptor may catch more information of the motion.

3.3 Comparison

Table 5 gives the comparison of performance with previous results on UT-Interaction dataset. All the numbers are copied from original articles. As shown in Table 5, our results outperform the previous ones. The accuracy of punching is the lowest because of the confusion with pushing. Our best result comes from "DT + MBH + SC_s $+\chi^2$ SVM" and "DT + combined + VQ/SC_s $+\chi^2$ SVM" on the set #1 and #2. The baseline experiments are "STIP + SVM" and "Cuboid + SVM" in [14]. Obviously, those sparse space-time volumes cannot deliver sufficient information for this interaction dataset. [17] used a Hough transform-based method to classify videos. A pedestrian detection algorithm was also adopted for their better classification. Particularly for the interaction challenge, they have modeled each actor's action using their voting method, forming a hierarchical system consisting of 2-levels. After observing all the dense trajectories in videos, we find that most of the DTs we used are densely catched at motion body. Our results benefit from the sufficient information at human body just as dense patches in image classification.

4. CONCLUSIONS

In this paper, we have explored the DT features and feature encoding approaches on human interaction data. The best performance up-to-date is achieved on UT-Interaction. Experiments demonstrate that sum or average pooling is superior to max pooling due to the large amount of DTs and variable foregrounds on UT-Interaction. The reason why VQ is better than other "localized" encoding methods is thoroughly analysed by a view of data manifold. Our future work will focus on how to accelerate the process of using dense features in videos.

5. ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (No. 61002042, 61071184, 60972111, 61036008), Shenzhen Basic Research Program (JC201005270350A, JCYJ20120903092050890, JCYJ20120617114614 438), 100 Talents Programme of Chinese Academy of Sciences, Guangdong Innovative Research Team Program (No.201001D0104648280), Research Funds for the Doctoral Program of Higher Education of China (No. 20100184120009), Program for Sichuan Provincial Science Fund for Distinguished Young Scholars (No. 2012JQ0029, No. 13QNJJ0149), the Fundamental Research Funds for the Central Universities (No. SWJTU09CX032, SWJTU10CX08) and the 2013 Doctoral Innovation Funds of Southwest Jiaotong University.

6. REFERENCES

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, volume 2, pages 1395–1402, 2005.

[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, pages 2559–2566, 2010.

[3] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, pages 778–785, 2011.

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *ECCV*, pages 428–441, 2006.

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[6] A. Klaser, M. Marszałek, C. Schmid, et al. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011.

[8] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008.

[10] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *NIPS*, volume 19, pages 801–808, 2007.

[11] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*, pages 2486–2493, 2011.

[12] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, pages 2929–2936, 2009.

[13] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, pages 1036–1043, 2011.

[14] M. Ryoo, C.-C. Chen, J. Aggarwal, and A. Roy-Chowdhury. An overview of contest on semantic description of human activities (sdha) 2010. *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 270–285, 2010.

[15] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, volume 3, pages 32–36, 2004.

[16] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *MM*, pages 357–360. ACM, 2007.

[17] D. Waltisberg, A. Yao, J. Gall, and L. Van Gool. Variations of a hough-voting action recognition system. *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 306–312, 2010.

[18] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

[19] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, Mar. 2013.

[20] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.

[21] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *ECCV*, pages 650–663, 2008.

[22] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801, 2009.

[23] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, pages 492–497, 2009.