A Temporal Bayesian Model for Classifying, Detecting and Localizing Activities in Video Sequences

Manavender R. Malgireddy
University at Buffalo
mrm42@buffalo.edu

Ifeoma Inwogu
University at Buffalo
inwogu@buffalo.edu

Venu Govindaraju
University at Buffalo
govind@buffalo.edu

Abstract

We present an framework to detect and localize activities in unconstrained real-life video sequences. This is a more challenging problem as it subsumes the activity classification problem and also requires us to work with unconstrained videos. To obtain real-life data, we have focused on using the Human Motion Database (HMDB), a collection of realistic video clips. The detection and localization paradigm we introduce uses a keyword model for detecting key activities or gestures in a video sequence. This process is analogous to the use of keyword or key-phrase detection in speech processing. The method learns models for the activities-of-interest during training, so that when presented with a network of activities (a representation of video sequences) at testing, the goal is to detect the keywords in the network. Our approach for classification outperformed all the current state-of-the-art classifiers when tested on two publicly available datasets, KTH and HMDB. We also tested this paradigm for spotting gestures via a one-shot-learning approach on the CHALEARN gesture dataset and obtained very promising results. Our approach was ranked amongst the top-5 best performing techniques in the CHALEARN 2012 gesture spotting competition.

1. Introduction

Activity recognition: given a sequence of images with one or more persons performing one or many different activities over time, can a system be designed to fully automatically recognize what activity is being performed in the sequence and in what specific frames it occurred? (definition adapted from Turuga *et al*. [21])

Till date, much of the computer vision community has approached this problem from a single activity perspective where the problem is reduced to classifying a sequence of images containing one activity. Hence given an image sequence, the assumption already exists that only one major activity from a known class of activities occurs in that sequence. In addressing this problem, other sub-problems that have emerged include i) investigating the effectiveness of different types of spatiotemporal features involved in classifying activities; (ii) unsupervised monitoring and anomaly detection of scenes such as traffic monitoring or crowded site monitoring; (iii) investigating the effectiveness of statistical models for classifying activities.

Although this problem in itself is still an open and challenging area of research, in our proposed work, we remove the simplifying assumption that only one activity type exists in an image sequence. Hence, given an image sequence containing many different activities, our goal is to determine whether or not one or more activities (from a known class of activities) occurs in that sequence, and to localize exactly where it occurs. This is a much more challenging and realistic problem than the current state-of-the-art approaches to activity recognition since it subsumes at least two of the current activity recognition sub-problems. In addition, we develop a framework for intelligently and effectively spotting known activities from a random sequence of activities (known and/or unknown).

The key challenges in activity spotting in general are two-fold (i) the search space for locating activities in unconstrained videos can be very large, increasing with increase in video length; and (ii) human actions/activities can exhibit tremendous variations within a single activity class, *i.e*. the same type of activity can be viewed differently because of variations in appearances, view points, lighting conditions, extent of motion, *etc*.

To overcome some of these challenges, we propose a new statistical model for activity classification and an efficient framework for spotting on video sequences containing zero or more of the activities learned by that model. We represent an activity as a temporal sequence of discretized motions, governed by the Markov assumption, and develop a hierarchical model where the discretized motions are modeled as distributions over observations. We evaluate the performances of both the classification and spotting tasks on publicly available datasets (KTH [19], HMDB [13]) and CHALEARN gesture dataset.

978-1-4673-1612-5/12/$31.00 ©2012 IEEE

2. Related Work

Features and interest points

Extensive research has gone into the study of the recognition of human activities in videos [1]. The study can be grouped into two main categories (i) spatio-temporal feature extraction; and (ii) model representation. Spatio-temporal interest points were earlier introduced by Laptev and Linde [14] and since then other interest-point-based detectors such as those based on spatio-temporal Hessian matrix [26] and Gabor filters [3, 6] have been proposed. Various other descriptors such as those based on HoG/HoF) [15], HoG3D [11], 3D-SIFT [20] and Local Trinary Patterns [28] have also been proposed to describe interest points. More recently descriptors based on tracking interest points have been explored [17, 16]; these use standard KLT trackers[1] to track interest points. In a recent paper by Wang *et al*. [23], the authors performed an evaluation of local spatio-temporal features for action recognition and showed that dense sampling of feature points significantly improved classification results compared to sparse interest points. Similar results were also shown by [18] for image classification.

Statistical models and activity spotting

Graphical models are extensively used for activity recognition. For example, Topic Modeling or Latent Dirichlet Allocation (LDA) [2] has been used to cluster low level features from videos into activities [25, 27, 24, 10]. The use of topic modeling in activity recognition aims to automatically discover "topics" which are co-occurrences of "visual words" or bags-of-activity-descriptors. These visual words are then clustered into "documents" or pre-defined activity classes. One drawback of strictly the LDA-based approaches is the absence of temporal modeling, *i.e*. restricting the modeling of the correlation of different motion patterns over time.

Gaur *et al*. [7] proposed a model based on a string representation of the video. The string represented the spatio-temporal ordering of interest points. Brendel *et al*. [4] proposed a model to represent videos by spatiotemporal graphs, where the nodes corresponded to multiscale video segments and edges captured hierarchical, temporal and spatial relationships.

Few methods have been proposed for activity spotting and among them include the work of Yuan *et al*. [29] who represented a video as a 3D volume and activities-of-interest as subvolumes. The task of activity spotting was therefore reduced to one of performing an optimal search for activities in the video. Another work in spotting was presented by Derpanis *et al*. [5] who introduced a local descriptor of video dynamics based on visual spacetime oriented energy measures. Similar to the previous work, their input was also a video which was search for a specific action. The limitation in these techniques is their inability to adapt to changes in view points, scale, appearance etc. Rather than being defined on the motions patterns involved in an activity, these methods performed a somewhat template matching type of technique and such methods do not readily generalize to new environments exhibiting a known activity. Both methods report their results on the KTH and CMU datasets.

[1] Kanade-Lucas-Tomasi Feature Tracker

3. Proposed Approach

In this section, we present an overview of our end-to-end framework describing how we train the model and then perform classification and spotting tasks. Figure 1 shows a high level process flow for how the temporal model is is learned. As shown, a probabilistic dynamic signature is created for every activity class to be learned and activity recognition becomes a problem of finding the most likely distribution to generate the test video. Activity spotting becomes a problem of decoding a network of combined signatures (as shown in Figure 3) to determine which component of the network could have generated a particular activity class. Additional details on the decoding process are provided in Section 3.2.

3.1. Generating a signature for an activity class

3.1.1 Computing the observables for the model

The first step in computing the probabilistic signature for an activity class involves the extraction of interest points from the frames (sampled from the videos at 30 fps) of the training videos containing only that activity. Because each frame can contain a large number of redundant points which have no bearing to the activity being learned, interest points are used for pruning. The specific methods for extracting interest points on the different datasets are provided in Section 4

Once the interest points are obtained, two key feature descriptors the Histogram of Gradients (HoG) and Histogram of Flow (HoF)[15] are implemented to obtain a dense population of descriptors over the entire space of activities. Using k-means clustering, the descriptors are converted to "visual words" or bags-of-activity-descriptors. If a frame is analogous to a document, the goal of the LDA process here is to to automatically discover the "topics" which are co-occurrences of the visual words. The visual words are thus re-expressed in terms of topics so that after LDA, each frame can be seen to be made up of visual words which belong to different topics.

Unlike the previous approaches where the clustering techniques such as LDA do away with the temporal aspect of the problem, our approach only uses LDA to reduce

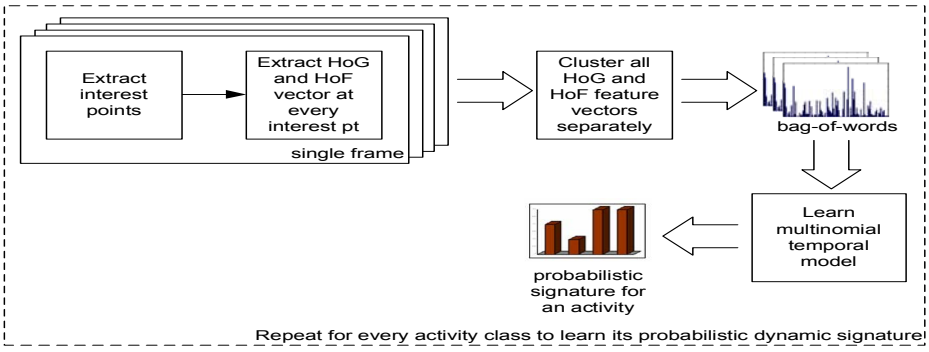

Figure 1. Process flow for the training cycle

the size and refine the meaning of our observable features. By explicitly applying a temporal model over these observables, we can still successfully model the correlation of different motion patterns over time. The observable feature for one frame is therefore the histogram over the visual words in that frame.

3.1.2 Computing the parameters for the model

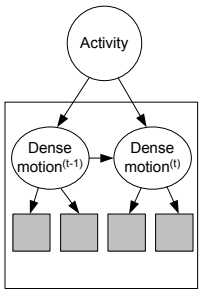

Figure 2. Graphical Model

In a general sense, our model can be interpreted as Hidden Markov Model (HMM) with states and observations but unlike classic HMM, this model has multiple channels/descriptors, where each channel is represented as a distribution over the visual words corresponding to that channel. In contrast to the classic HMM, our model can have multiple observations per state and channel. We refer to this model as mcHMM (multiple channel HMM). Figure 2 shows a graphical representation of the mcHMM. We choose to separate the HoG channel from the HoF channel because they capture very different spatiotemporal properties from the video input. In our mcHMM, we refer to the states of the HMM as dense motions.

Hence, to determine the probabilistic signature of an activity class, one mcHMM is trained for each activity. The generative process for mcHMM can be given as follows: (i) dense motion is sampled from an activity based on the transition matrix for that activity; (ii) a frame-feature (comprising of the distribution of visual words) is sampled according to a multinomial distribution for that dense motion[2]. 3) Repeat this for each frame. Similar to a classic HMM, the parameters for the mcHMM are therefore:

1. Initial state distribution $\pi = \{\pi_i\}$
2. State transition probability distribution $A = \{a_{ij}\}$
3. Observation densities for each state and descriptor $B = \{b_i^d\}$

The joint probability distribution of observations (O) and hidden state sequence (Q) given the parameters of the multinomial representing a hidden state (λ) can be expressed as:

$$P(O, Q|\lambda) = \pi_{q_1} b_{q_1}(O_1) \prod_{t=2}^{T} a_{q_{t-1}q_t} \cdot b_{q_t}(O_t) \quad (1)$$

where $b_{q_t}(O_t)$ is modeled as follows

$$\begin{aligned} b_{q_t}(O_t) = & \prod_{d=1}^{D} b_q^d(O_t^d) \\ = & \prod_{d=1}^{D} Mult(O_t^d|b_q^d) \end{aligned} \quad (2)$$

and D is the number of descriptors.

Expectation-Maximization (EM) is implemented to find the maximum likelihood estimates. The update equations for the model parameters are:

$$\hat{\pi} = \sum_{r=1}^{R} \gamma_1^r(i); \quad (3)$$

$$\hat{a}_{ij} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T} \eta_t^r(i,j)}{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t^r(i)} \quad (4)$$

$$\hat{b}_j^d(k) = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t^r(j) \cdot \frac{n_t^{d,k}}{n_t^{d,\cdot}}}{\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t^r(j)} \quad (5)$$

where R is number of videos and $\gamma_1(i)$ is the expected number of times the activity being modeled started with dense motion i;

$\eta_t^r(i,j)$ is the expected number of transitions from dense motion i to dense motion j and $\gamma_t^r(i)$ is the expected number of transitions from dense motion i;

$n_t^{d,k}$ is the number of times that visual word k occurred in descriptor d at time t and $n_t^{d,\cdot}$ is the total number of visual words that occurred in descriptor d at time t.

[2]dense motions are modeled as multinomials since our input observables are discrete

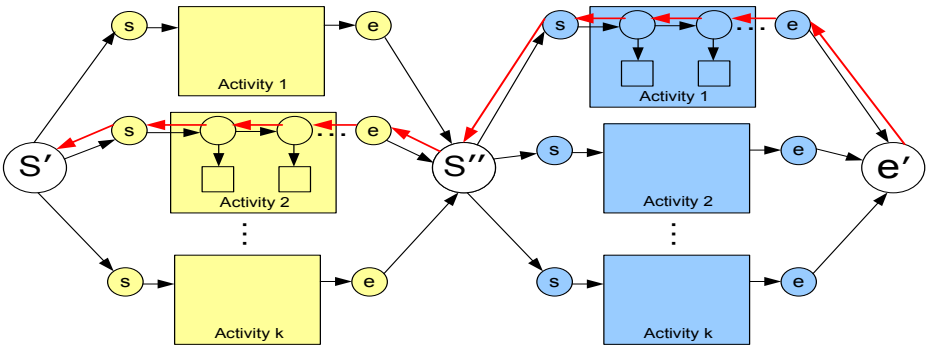

Figure 3. Activity spotting by computing likelihoods via Viterbi decoding. The toy example shown assumes there are only two activities in any test video, where the first activity is from the yellow set, followed by one from the blue set. The image also shows an example of a putative decision path in red, after the decoding is completed (image best viewed in color)

3.2. Activity recognition and activity (gesture) spotting

The activity recognition problem is thus reduced to an inference problem where given a new previously unseen test video, and the model parameters or probabilistic signatures of known activity classes, the goal is establish which activity class distribution most likely generated the test video. This type of inference can be achieved using the Viterbi algorithm.

Figure 3 shows an example of the stacked mcHMMs involved the activity spotting task. Our current method of activity spotting using the probabilistic signature for an activity class requires that the number of activities in the test videos are the same and are known in advance. So in the toy example shown, a test video is comprised of two activities only. (This design choice was driven by our participation in the Chalearn competition where the number of activities in every test video was given to be 5). Although this is currently a limitation, the process can extended by implementing a more general infinite network structure where in figure 3, there would be a feedback arrow from the final state e' to the intermittent start state s''.

Again, the Viterbi decoding algorithm is implemented to traverse the stacked network. Putative decisions arise when the Viterbi path crosses the keyword portion of the model. The ratio between the likelihood of the Viterbi path that passed through the keyword model and the likelihood of an alternate path that passes solely through the non-keyword portion is then used to score the occurrence of a keyword, where a keyword here refers to an activity class. An empirically chosen threshold value is used to select the occurrence of a keyword in the signal being decoded.

4. Implementation

In this section, we describe the details of implementing our proposed approach on several publicly available benchmark datasets.

4.1. Interest points and features

Depending on the task as hand, the source and size of the training set can vary extensively. For example, for performing the *classification task* on KTH, the dataset comprised of about 2300 video clips used to train/test six activities. For HMDB, a significantly more complex and diverse dataset, there were 7000 video clips to train/test 51 complex activities. For the Chalearns dataset, since the task-at-hand was spotting via one-shot learning, only one video per class is provided to train an activity (or gesture as the dataset organizers referred to them).

Interest points were obtained in the KTH and HMDB dataset in a similar manner by sampling dense points in every frame in the video and then track these points for the next L frames. For each of these tracks, motion boundary histogram descriptors based on the HoG and HoF descriptors were extracted. Because the HMDB dataset comprises of real-life scenes which contain people and activities occurring at multiple scales, the frame-size in the video was reduced by a factor of 2 repeatedly and motion boundary descriptors were extracted at multiple scales.

Figure 4. Interest points (shown in red) for 2 consecutive frames from the Chalearn dataset

In the Chalearn dataset, since the videos were comprised of RGB-D images, we extracted interest points by taking the difference between two consecutive depth images. Figure shows an example of two consecutive depth images from the dataset, with the interest points superimposed. HoG and

HoF descriptors were then extracted at each interest point so that similar descriptors could be obtained as those from the KTH and HMDB datasets. For HoG and HoF implementation, we used a spatial size of 32x32 and 4 cells of size 16x16.

Next, the feature descriptors were clustered to obtain the first round of visual words. In general, from the literature, in order to limit complexity, researchers randomly select a finite number (roughly in the order of 100,000) of training features which could prove reasonable for sparse features and small datasets. In dealing with dense features, the amount of descriptors generated at multiple frames was significantly more. For example, if there are 2000 videos for one activity and for each video there are about 15,000 feature descriptors on average, resulting in a total of 30,000,000 features (for one scale only). Randomly selecting say 100,000 or 0.3% of the data as the dimension of the resulting visual words would not have sufficiently represented such a large space.

To attenuate this issue, we divide the construction of visual words into a two step process where we first construct visual words for each activity class separately in parallel. Then we use the visual words obtained for each class as the input samples to cluster the final visual words. This process significantly reduces the amount of data to be eventually presented to the model.

5. Experiments and Results

In this section we present (i) the results obtained from performing activity classification on video sequences created from the Human Motion Database (HMDB) [13] and KTH database; and (ii) we also present results of gesture spotting from the CHALEARN gesture dataset [8]. The CHALEARN gesture dataset has both depth and RGB videos, the task is to perform one short learning, which is the main challenge of this dataset. The HMDB is currently the most realistic test video database for activity recognition, while KTH is somewhat representative of the ceiling dataset in activity recognition (more recent papers achieve close to 100% activity classification due to the simplicity of the data). By testing on the two extreme benchmark datasets, we show that our classification framework competes well with the current state-of-the-art activity classification techniques.

5.1. Activity Classification Results

In order to compare our framework to the other current state-of-the-art methods, we ran the standard activity classification test on both the KTH and HMDB test sequences. Table 1 shows the comparison of accuracies obtained. Our best accuracy on the KTH dataset was obtained using 2000 visual words and 25 states.

Method	Accuracy
Laptev *et al*. [15]	91.8%
Yuan *et al*. [29]	93.3%
Wang *et al*. [22]	94.2%
Gilbert *et al*. [9]	94.5%
Kovashka and Grauman [12]	94.53%
Our Method	**94.67**

Table 1. Comparison of our proposed model and features with state-of-the-art activity classification methods for KTH dataset

Similarly, we performed activity classification tests on all 51 categories of the HMDB and the results are shown in Table 2. We outperform the only currently reported accuracy results on this dataset. Rows 1-3 show accuracies for testing performed on the stabilized version of the data. When tested on 10 activities stableand unstable, the results were 66.67 % and 57.67 % respectively. Our best accuracy on the HMDB was obtained using 1000 visual words and 25 states.

Method	Accuracy
Best results on 51 activities by Kuehne *et al*. [13]	23.18%
Best results on 10 activities by Kuehne *et al*. [13]	54.3%
Our best results on 51 activities	**25.64%**
Our best results on 10 activities	**66.67%**
Our best results on 10 activities (original)	57.67%

Table 2. Comparison of our proposed model and features with state-of-the-art activity classification methods for HMDB dataset

5.2. One-Shot-Learning using Chalearn dataset

Details on the dataset and how it can be used can be found in [8]. We performed one-shot-learning using our proposed model where a model was learned for every gesture (every training video). For gesture spotting, a network of gestures was created by connecting models of each gestures learned. We used edit distance to evaluate our performance. Table 3 shows the results of one-shot-learning. We placed fourth in the final results of the Chalearn 2012 gesture challenge using this method.

Method	Dataset	edit distance
Ours	Development	0.26336
Ours + LDA	Development	0.2409
Ours	Validation	0.26036
Ours + LDA	Validation	**0.23328**
baseline	Validation	0.59978
Top Ranking Participant	Validation	0.1426

Table 3. Results for Chalearn gesture dataset

6. Conclusion and future work

We have presented an end-to-end temporal Bayesian framework for activity classification and spotting. Our framework competes well with the current state-of-the-art techniques in activity classification and verify this by testing on the two extreme datasets the trivial KTH, and the very complex realistic HMDB. We also show the efficacy of our model by participating in Chalearn gesture challenge and finished top-5 in the competition.

In the future, we intend to perform a more rigorous study of the proposed method of activity spotting by varying the threshold score at which a keyword is acknowledged. An ROC curve can be developed in order to select optimal thresholds values.

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43:16:1–16:43, Apr. 2011. 2

[2] D. M. Blei and J. D. Mcauliffe. Supervised Topic Models. In *NIPS*, 2007. 2

[3] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1948–1955, 2009. 2

[4] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, 2011. 2

[5] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 1990–1997, 2010. 2

[6] P. Dollr, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *In VS-PETS*, pages 65–72, 2005. 2

[7] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A "string of feature graphs" model for recognition of complex activities in natural videos. In *ICCV*, 2011. 2

[8] C. Gestures. Chalearn gesture dataset (cgd2011), chalearn, 2011. 5

[9] A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):883–897, 2011. 5

[10] T. Hospedales, S.-G. Gong, and T. Xiang. A Markov Clustering Topic Model for Mining Behaviour in Video. In *ICCV*, pages 1165–1172, 2009. 2

[11] A. KlŁser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *In BMVC*, 2008. 2

[12] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, pages 2046–2053, 2010. 5

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 1, 5

[14] I. Laptev and T. Lindeberg. Space-time interest points. In *IN ICCV*, pages 432–439, 2003. 2

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions From Movies. In *CVPR*, pages 1–8, 2008. 2, 5

[16] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features, 2009. 2

[17] R. Messing, C. Pal, and H. Kautz. Activity Recognition Using the Velocity Histories of Tracked Keypoints. In *ICCV*, 2009. 2

[18] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *In Proc. ECCV*, pages 490–503. Springer, 2006. 2

[19] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR (3)*, pages 32–36, 2004. 1

[20] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 357–360, New York, NY, USA, 2007. ACM. 2

[21] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine Recognition of Human Activities: A Survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, Sept. 2008. 1

[22] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Jun 2011. 5

[23] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, sep 2009. 2

[24] X. Wang, X. Ma, and W. Grimson. Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models. *IEEE TPattern Anal. Mach. Intell.*, 31:539–555, 2009. 2

[25] Y. Wang and G. Mori. Human Action Recognition by Semilatent Topic Models. *IEEE TPattern Anal. Mach. Intell.*, 31:1762–1774, 2009. 2

[26] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag. 2

[27] T. Xiang and S. Gong. Video Behavior Profiling for Anomaly Detection. *IEEE TPattern Anal. Mach. Intell.*, 30:893–908, May 2008. 2

[28] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009. 2

[29] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, 2009. 2, 5