MoSIFT: Recognizing Human Actions in Surveillance Videos

CMU-CS-09-161

Ming-yu Chen and Alex Hauptmann

School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213

September 24, 2009

Abstract

The goal of this paper is to build robust human action recognition for real world surveillance videos. Local spatio-temporal features around interest points provide compact but descriptive representations for video analysis and motion recognition. Current approaches tend to extend spatial descriptions by adding a temporal component for the appearance descriptor, which only implicitly captures motion information. We propose an algorithm called MoSIFT, which detects interest points and encodes not only their local appearance but also explicitly models local motion. The idea is to detect distinctive local features through local appearance and motion. We construct MoSIFT feature descriptors in the spirit of the well-known SIFT descriptors to be robust to small deformations through grid aggregation. We also introduce a bigram model to construct a correlation between local features to capture the more global structure of actions. The method advances the state of the art result on the KTH dataset to an accuracy of 95.8%. We also applied our approach to 100 hours of surveillance data as part of the TRECVID Event Detection task with very promising results on recognizing human actions in the real world surveillance videos.

1. Introduction

Visual surveillance systems collect huge amounts of video but a human must still review most of the data to extract informative knowledge. Our goal is to enable automatic recognition of many different types of behavior, without careful labeling, over large archives. Once events and behaviors are recognized, the results are used for retrieval, alerting and summarization of the data.

Methods based on feature descriptors around local interest points are now widely used in object recognition. This part-based approach [16] assumes that a collection of distinctive parts can effectively describe the whole object. Compared to global appearance descriptions, a part-based approach has better tolerance to posture, illumination, occlusion, deformation and cluttered background. Recently, spatio-temporal local features have been used for motion recognition in video. The spatio-temporal features share some of the same desirable properties as local features of images and also achieve good performance on human action recognition tasks [1, 2, 3, 4, 5, 6, 7].

The key to the success of part-based methods is that the interest points are distinctive and descriptive. Therefore, interest point detection algorithms play an important role in a part-based approach. The straightforward way to detect a spatio-temporal interest point is to extend a 2D interest point detection algorithm. Laptev et al. [1] extended 2D Harris corner detectors to a 3D Harris corner detector, which detects points with high intensity variations in both spatial and temporal dimensions. On other words, a 3D Harris detector finds spatial corners with velocity change. This detector can produce compact and distinctive interest points. However, since the assumption of change in all 3 dimensions is quite restrictive, very few points results and many motion types may not be well distinguished. Dollar et al. [4] discarded spatial constraints and focused only on the temporal domain. Since they relaxed the spatial constraints, their detector detects more interest points than a 3D Harris detector by applying Gabor filters on the temporal dimension to detect periodic frequency components. Although they state that regions with strong periodic responses normally contain distinguishing characteristics, it is not clear that periodic movements are sufficient to describe complex actions. Oikonomopoulos et al. [7] computed the entropy of small spatio-temporal volumes and select volumes with large entropy as interest points.

The above detection methods find informative but fairly sparse interest points among video cuboids. In contrasts, the highly successful Scale-Invariant Feature Transform (SIFT) for object recognition detects many interest points in an image and descriptors of these points are used to match static objects. Since recognizing human motion is more complicated than object recognition, motion recognition is likely to require with enhanced local features that provide both shape and motion information.

The goal of this paper is to achieve robust human action recognition in the TREC Video Retrieval Evaluation (TRECVID 2008) [8] real-world London Gatwick airport surveillance videos. We first re-examine spatio-temporal feature detection and description for improved recognition performance. We observed that humans can distinguish actions from the appearance of feature points and their movements. Therefore, we propose an algorithm called MoSIFT, inspired by the Scale Invariant Feature Transform (SIFT) [9] but developed for video interest points and features instead of image points/features. Our main insight is to treat spatial dimensions and temporal dimension separately. We nominate candidate points by detecting distinctive appearances and extract spatio-temporal local features if the candidate points contain movement. This detection algorithm extracts a good number of interest points. A MoSIFT descriptor is designed to represent the feature point in two parts: The first is an aggregated histogram of gradients (HoG) to describe the spatial appearance. The second part is an aggregated histogram of optical flow (HoF) which indicates the movement of the feature point. The aggregation of histograms provides better invariance to any deformation. We also propose a bigram model to capture more global shape and sequence information. The bigram can be defined as a pair-wise constraint and is auxiliary to a unigram bag-of-word feature in any local feature approach.

We describe the MoSIFT algorithm for interest point detection and feature description in section 2. Section 3 discusses our pair-wise constraint model, which is derived from bigram models in text classification. In section 4, we present experimental performance results on the KTH dataset [2] and analyze the very large TRECVID London Gatwick surveillance dataset as an example of a real world application. We conclude with future work and a summary.

2. MoSIFT

This section presents our MoSIFT algorithm to detect and describe spatio-temporal interest points. In part-based methods, there are three major steps: detecting interest points, constructing a feature descriptor, and building a classifier. Detecting interest points reduces the whole video from a volume of pixels to compact but descriptive interest points. Therefore, we desire to develop a detection method, which detects a sufficient number of interest points containing the necessary information to recognize a human action. The MoSIFT algorithm detects spatially distinctive interest points with substantial motions. We first apply the well-know SIFT algorithm to find visually distinctive components in the spatial domain and detect spatio-temporal interest points with (temporal) motion constraints. The motion constraint consists of a 'sufficient' amount of optical flow around the distinctive points. Details of our algorithm are described in the following sections.

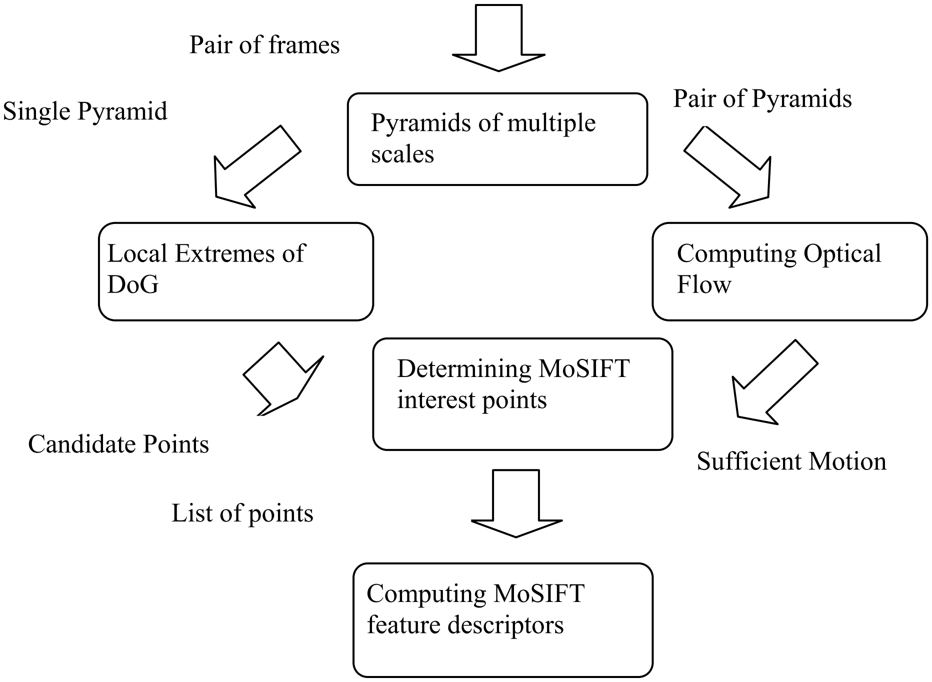


Figure 2: System flow graph of the MoSIFT algorithm. A pair of frames is the input. Local extremes of DoG and optical flow determine the MoSIFT points for which features are described.

2.1 MoSIFT interest point detection

Figure 2 demonstrates our MoSIFT algorithm. The algorithm takes a pair of video frames to find spatio-temporal interest points at multiple scales. Two major computations are applied: SIFT point detection and optical flow computation according to the scale of the SIFT points.

SIFT interest points are scale invariant and all scales of an image must be considered. Lowe [9] used a Gaussian function as a scale-space kernel to produce a scale space of the image. The whole scale space is divided into a sequence of octaves and each octave is divided into a sequence of intervals, where each interval is a scaled frame. The number of octaves and intervals is determined by the frame size. The size relationship between two adjacent octaves is in powers of 2. The first interval in the first octave is the original frame. In each octave, the first interval is denoted as $I(x,y)$. We can denote each interval as

$$L(x,y,k\delta)=G(x,y,k\delta)*I(x,y) \qquad (1)$$

where $*$ is the convolution operation in x and y, and $G(x,y,k\delta)$ is a Gaussian smoothing function. Difference of Gaussian (DoG) images are then computed by subtracting adjacent intervals

$$D(x, y, k\delta) = L(x, y, k\delta) - L(x, y, (k-1)\delta) \quad (2)$$

Once the pyramid of DoG images has been obtained, the local extremes (minima/maxima) of the DoG images across adjacent scales are used as the interest points. This is done by comparing each pixel in the DoG images to its eight neighbors at the same interval and nine corresponding neighboring pixels in each of the neighboring intervals. The algorithm scans through each octave and interval in the DoG pyramid and detects all possible interest points at different scales.

However, SIFT is designed to detect distinctive interest points in a still image. The candidate points are distinctive in appearance, but they are independent of the motions or actions in video. For example, a cluttered background can produce many interest points unrelated to human actions. Clearly, only interest points with sufficient motion will provide the necessary information for action recognition. The widely used optical flow approach detects the movement of a region by calculating where a region moves in the image space by measuring temporal differences. Compared to video cuboids or volumes, optical flow explicitly captures the magnitude and direction of a motion, instead of implicitly modeling motion through appearance change over time. Our belief is that explicit motion measurement is essential for recognizing actions.

In the interest point detection part of the MoSIFT algorithm, optical flow pyramids are constructed over two Gaussian pyramids. Multiple-scale optical flows are calculated according to the SIFT scales. A local extreme from DoG pyramids can only become an interest point if it has sufficient motion in the optical flow pyramid. We assume that a complicated action can be represented by the combination of a reasonable number of interest points. Therefore, we do not assign strong constraints to spatio-temporal interest points. As long as a candidate interest point contains a minimal amount of movement, the algorithm will extract this point as a MoSIFT interest point. MoSIFT interest points are scale invariant in the spatial domain. However, they are not scale invariant in the temporal domain. Temporal scale invariance could be achieved by calculating optical flow on multiple scales in time. However, we want to select distinctive interest points with sufficient motion where humans could 'see' the action based on these points and machines could learn an action model. Therefore, a small motion is sufficient at each interest point rather than imposing a complex motion constraint. Ultimately, this is still an open research topic.

2.2 MoSIFT feature description

In most current work on action recognition, much emphasis is placed on interest point detection and action model learning. However, the feature descriptor is an important step which is almost ignored. Most work [1,2,4,5] uses histograms of gradients to describe the appearance of interest volumes or cuboids. Some recent work [10,11] includes histograms of optical flow to boost performance.

Since MoSIFT point detection is based on DoG and optical flow, it is natural that our descriptor leverages these two features. Instead of combining a complete HoF classifier

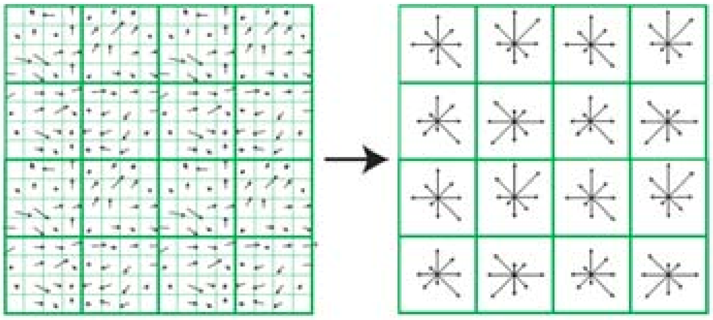


Figure 3: Grid aggregation for SIFT/MoSIFT feature descriptors. Pixels in a neighborhood are grouped into 4x4 regions. An orientation histogram with 8 bins is formed for each region resulting in a 128 element vector. MoSIFT concatenates aggregated grids for both appearance and motion for a 256 element descriptors vector.

with a complete HoG classifier, we build a single feature descriptor, which concatenates both HoG and HoF into one vector, which is also called 'early fusion'. We believe appearance and motion information together are the essential components for a classifier. Since an action is only as represented by a set of spatio-temporal point descriptors, the descriptor features critically determines the information used by later recognition steps.

It is often underappreciated that the original SIFT descriptor captures local appearance with an aggregated histogram of gradients from neighboring regions. This gives the SIFT descriptor better tolerance to partial occlusion and deformation. When an interest point is detected, a dominant orientation is calculated and all gradients in the neighborhood are rotated according to the dominant orientation to achieve rotation invariance. The magnitude and direction for the gradient are calculated for every pixel in a region around the interest point in the Gaussian-blurred image L. An orientation histogram with 8 bins is formed, with each bin covering 45 degrees. Each sample in the neighboring window is added to a histogram bin and weighted by its gradient magnitude and its distance from the interest point. Pixels in the neighboring region are normalized into 256 (16x16) elements. Elements are grouped as 16 (4x4) grids around the interest point. Each grid has its own orientation histogram to describe sub-region orientation. This leads to a SIFT feature vector with 128 dimensions (4x4x8 = 128). Each vector is normalized to enhance invariance to changes in illumination. Figure 3 illustrates the SIFT descriptor grid aggregation idea.

MoSIFT adapts the idea of grid aggregation in SIFT to describe motions. Optical flow detects the magnitude and direction of a movement. Thus, optical flow has the same properties as appearance gradients. The same aggregation can be applied to optical flow in the neighborhood of interest points to increase robustness to occlusion and deformation. The main difference to appearance description is in the dominant orientation. Rotation invariance is important to appearance since it provides a standard to measure the similarity of two interest points. In surveillance video, rotation invariance of appearance remains important due to varying view angles and deformations. Since surveillance video is captured by a stationary camera, the direction of movement is actually an important (non-invariant) vector to help recognize an action. Therefore, we omit adjusting for orientation invariance in the MoSIFT motion descriptors. The two aggregated histograms (appearance and optical flow) are combined into the MoSIFT descriptor, which now has 256 dimensions.

3. Bigram model of video code words

The bag-of-words feature representation is often used to represent a motion event using spatio-temporal interest points. A video codebook is constructed by clustering spatio-temporal interest points. Each interest point is then assigned to its closest vocabulary word (a cluster) and the histogram of video words is computed over a space-time volume to describe an action.

A bag-of-words feature representation is easy to compute and efficient for describing an action. However, its histogram does not contain any spatial and temporal constraints, which leads to loss of shape and periodicity information. In text analysis, a bigram model is often used to capture the co-occurrence of adjacent words in order to boost classification results [13]. We can apply this idea to video code words. Although it is computationally intractable to model all possible sequences of video code words in a space-time volume, co-occurrence of only two video words requires minimal computation and provides some spatial and temporal constraints that help model shape and periodic motion.

We first define adjacent video words as a pair of video words which co-occur in a kernel where d_s and d_t denote the spatial and temporal boundary. Experience has shown that good vocabulary sizes for action recognition are in the range of a hundred to a thousand words. Pair-wise correlations can result in very large numbers of pairs. Some research [14,15] reduces the number of correlations by clustering. Instead, we select bigrams based on their tf-idf weight (term frequency-inverse document frequency) which is common in information retrieval and text classification. Term frequency (tf) is the frequency of a bigram in the dataset. Inverted document frequency (idf) indicates how informative a bigram is by dividing the number of all actions by the number of actions containing this bigram, and then taking the logarithm of the quotient. All bigrams can then be ranked by their tf-idf weights and we pick a sufficient number of bigrams to provide extra constraints to enrich the bag-of-word features and boost action classification performance.

As we pick n bigrams with video codebook of m vocabularies, the histogram size will be $n+m$. We calculate the histogram as a vector:

$$\hat{H}(i) = \frac{1}{|p_i|} \sum_{p \in \{p_i\}}^{|p_i|} \frac{1}{|C|} \sum_{c \in C}^{|C|} h(p,c) \qquad (3)$$

$$h(p,c) = \exp(-g \times dis(p,c)) \qquad (4)$$

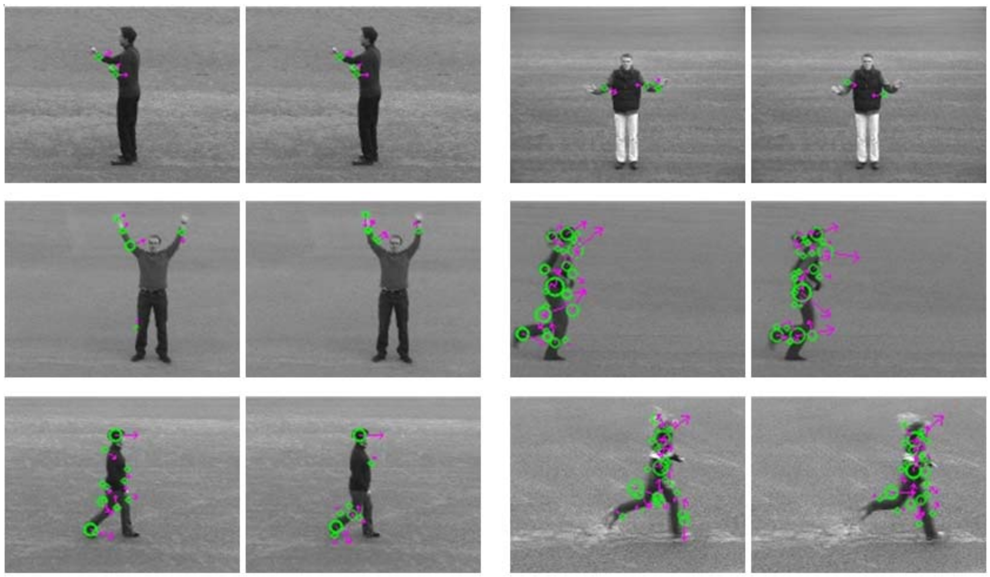


Figure 4: Some examples of MoSIFT from KTH dataset. In left two columns, from top to bottom are boxing, handwaving and walking. In right two columns, from top to bottom are handclapping, jogging and running. Green circle indicates interest points and purple arrow shows the direct to move. From image sequence, jogging and running are very similar.

where $\{p_i\}$ is the set of interest points with vocabulary label i and $|p_i|$ is the size of this vocabulary. C is the set of interest points around interest point p and $h(p,c)$ is a weighting function for a pair of interest points. If the pair is far apart, it contributes less to the histogram. g is a fixed parameter of $h(p,c)$ and $dis(p,c)$ measures the distance between interest points, a Euclidean distance in our case.

4. Experimental results

We first evaluate our algorithm on the KTH human motion dataset. The goal is to compare the performance to existing methods on a common standard. We then apply the framework to the London Gatwick airport surveillance dataset which consists of 100 hours of real world surveillance video by five cameras.

K-mean clustering is used to construct the video codebook. For classification, we use a support vector machine (SVM) with x^2 kernel. The x^2 kernel is defined as:

$$K(x_i, x_j) = \exp\left(-\frac{1}{A} D(x_i, x_j)\right) \qquad (5)$$

Method	Accuracy
MoSIFT with Bigram	95.83%
MoSIFT	95.0%
MoSIFT detection with HoG & HoF & Bigram	93.33%
MoSIFT detection with HoG & HoF	89.15%
MoSIFT detection with HoF	86.10%
MoSIFT detection with 3D HoG	84.28%
Temporal Gabor filter with 3D HoG (from [4])	81.50%
3D Harris corner with 3D HoG (from [2])	71.72%

Table 1: Comparison of different methods using the KTH dataset. HoG and HoF here indicate histograms without grid aggregation. 3D HoG indicates the histogram is calculated from a space-time volume.

method	Accuracy
MoSIFT	95.83%
Laptev et al. [10]	91.8%
Wong et al. [6]	86.62%
Nieble et al. [5]	83.33%
Dollár et al. [4]	81.50%
Schuldt et al. [2]	71.72%
Ke et al. [3]	62.96%

Table 2: Comparison of different methods using KTH

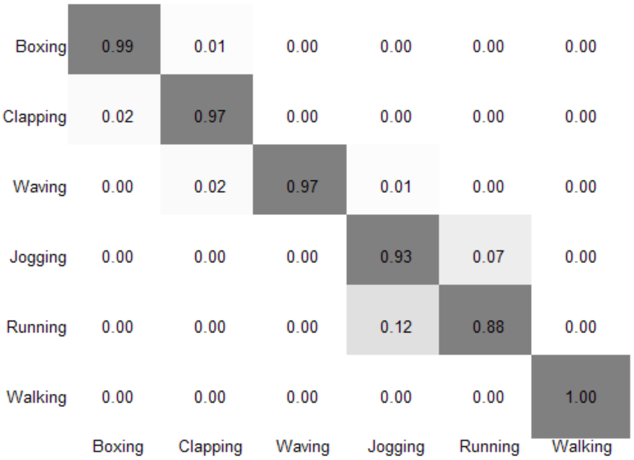
where A is a scaling parameter that can be determined though cross-validation. $D(x_i, x_j)$ is the x^2 distance defined as:

$$D(x_i, x_j) = \frac{1}{2}\sum_{i=1}^{m}\frac{(u_i - w_i)^2}{u_i + w_i} \qquad (6)$$

with $x_i = (u_1, \ldots, u_m)$ and $x_j = (w_1, \ldots, w_m)$. Prior work has shown that the x^2 kernel gives better performance with histogram features such as the bag of words [12]. Since we have multiple actions in both datasets, we adopted a one-vs-all strategy in the SVM classifiers.

4.1 Action recognition using the KTH dataset

The KTH human motion dataset is frequently used to evaluate event detection and recognition. It is also the largest widely available video dataset of human actions for



	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	0.99	0.01	0.00	0.00	0.00	0.00
Clapping	0.02	0.97	0.00	0.00	0.00	0.00
Waving	0.00	0.02	0.97	0.01	0.00	0.00
Jogging	0.00	0.00	0.00	0.93	0.07	0.00
Running	0.00	0.00	0.00	0.12	0.88	0.00
Walking	0.00	0.00	0.00	0.00	0.00	1.00

Figure 5: Confusion matrix of MoSIFT with bigrams for the KTH data. Jogging and running are the two most confusable actions.

researchers to evaluate and compare. The dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed by 25 different persons. Each person performs the same action four times under four different scenarios (outdoors, outdoors at a different scale, outdoors with camera moving and indoors). The whole dataset contains 598 video clips and each video clip contains only one action. KTH provides a common benchmark to evaluate and compare algorithms.

We extracted around 1.6 million interest points from the whole KTH dataset with the MoSIFT detector. Since the number of interest points is large, we randomly sampled 1% of the interest points to construct a video codebook. We chose a vocabulary size of 600 words for our codebook. Earlier experiments had shown that the performance is stable with codebook sizes above 500 words. We follow [5] in performing leave-one-out cross-validation to evaluate our approach. Leave-one-out cross-validation uses 24 subjects to build action models and then tests on the remaining subject. Performance is reported as the average accuracy of 25 runs. Figure 4 shows some MoSIFT detection examples.

We first wanted to evaluate the utility of our MoSIFT interest point detector by comparing against the well-known temporal Gabor filter [4] and the 3D Harris corner detector [1]. To make this comparison fair, the 3D HoG descriptor is used in all three detectors. When we indicate HoG and HoF in Table 1, we specifically mean histograms which contain the whole region of interest without aggregating grids. 3D HoG indicates that the histogram of gradients is calculated from space-time cuboids. The bottom three lines in Table 1 show that the MoSIFT detector extracts sufficient interest points to recognize human actions, slightly outperforming the Gabor and 3D Harris methods.

We then evaluated the efficiency of different descriptors. The MoSIFT detector was used with four different descriptors: 3D HoG, HOF, HoG & HoF, and grid aggregated HoG & HoF. The true MoSIFT descriptor is defined as the latter, i.e. grid-aggregated HoG and HoF. As mentioned in section 2, 3D HoG is often used and encodes motion by spatio-temporal appearance. In this evaluation, we compare explicit motion descriptors against implicit motion descriptors. Grid aggregated histogram descriptors theoretically should have better tolerance to occlusion, deformation and small shifts. We also want to illustrate that this comprehensive descriptor can substantially improve action recognition performance. Table 1 shows that 3D HoG representation results in an accuracy of 84.28% which is clearly better than the other interest point detectors. The performance gets another boost of 5% (to 89.15%) when we switch from 3D HoG to the combination of HoG and HoF. Note that in the HoG descriptor we only extract 2D HoG, which mainly represents shape. By adding grid aggregation, we improve the results another 6% (95.00%), which to our knowledge is the highest published result using a part-based method. The results support our notion that the descriptor is very important and provides a critical bridge from the video content to the feature space for machine learning.

The last evaluation on the KTH dataset relates to global information. In section 3, we obtained pair-wise constraints to enrich local features with shape and time sequence information by using a bigram model. We added bigrams our bag-of –word features in two different ways: the MoSIFT detector with non-aggregated HoG & HOF and the MoSIFT detector with full MoSIFT descriptor (aggregated HoG & HOF). The size of the kernel is 5x5x60, which is 5 pixels in the spatial dimensions and 60 frames in the temporal dimension. The number of bigrams we used was 300, which was determined to by reasonable through cross-validation. In fact, cross-validation shows that the first 300 bigrams significantly improve recognition performance. Beyond that, performance initially remains stable and eventually declines slightly as the number of bigram increases further. Table 1 shows that the bigram model improves the weaker descriptor by a substantial amount from 89.15% to 93.33%. However, it provides only a small improvement over the MoSIFT descriptor (95% to 95.83%). The high accuracy of the MoSIFT detector and descriptor at 95% means that among the 24 actions a subject performs, only 1 action is misrecognized. For certain actions in KTH such as running v.s. jogging, we found that even humans have difficulties in distinguishing them. Figure 5 shows the confusion table from our best result (MoSIFT with bigrams).

4.2 Event recognition using the Gatwick dataset

The 2008 TRECVID surveillance event detection dataset was obtained from London Gatwick International Airport. It consists of 50-hours (5 days x 2 hours/day x 5 cameras) of video in the development set and 49-hours in the evaluation set. There are about 190K frames per 2-hour video with an image resolution 720 x 576. This dataset contains highly crowded scenes, severely cluttered background, large variance in viewpoints, and very different performances of the same actions; all embedded in a huge amount of data. Together, these characteristics make action detection on this dataset a formidable challenge. To the best of our knowledge, human action detection on such a large, challenging task with these practical concerns has not been evaluated and reported prior to TRECVID 2008. In this evaluation, 10 events are evaluated: **ObjectPut, PeopleMeet, PeopleSplitUp, Pointing, CellToEar, Embrace, PersonRuns, ElevatorNoEntry, TakePicture,** and **OpposingFlow**. Standardized annotations of actions in the development set were provided by NIST.

Although this is ultimately a detection task ("find any event X"), we can also evaluate recognition ("what is this event") performance using the annotations provided by NIST.

Action	Chance	IG + 3D HoG	MoSIFT	MoSIFT Bigram
CellToEar	0.07	0.22	0.21	0.21
Embrace	0.08	0.25	0.28	0.29
ObjectPut	0.18	0.33	0.43	0.45
PeopleMeet	0.22	0.33	0.41	0.44
PeopleSplitUp	0.14	0.42	0.55	0.57
Pointing	0.26	0.38	0.41	0.41
PersonRuns	0.05	0.22	0.36	0.40
Average	0.14	0.31	0.38	0.40

Table 3: Comparison of different methods using the Gatwick dataset. We compare results using average precision (AP) as a metric. Chance denotes a random baseline. IG+HoG signifies that interest points were detected as high intensity gradients and HoG was used to describe these interest points [18]. MoSIFT indicates MoSIFT point detection and feature description without bigrams. MoSIFT Bigram indicates MoSIFT point detection and feature description with bigrams.

There were a total of 6439 events in the development set. Due to the size of the dataset, we detected and extracted MoSIFT descriptors only every five frames. The size of the video codebook was increased to 1000 after cross validation on the development set. Since the data were captured over 5 different days, we used 5-fold cross validation to evaluate our performance (see Table 3). There are not enough annotated examples for **OpposingFlow, ElevatorNoEntry** and **TakePicture** to run cross validation. Therefore, we do not report performance of these three tasks. We use average precision (AP) as the metric, which is typical for TRECVD high-level feature recognition. We first extracted interest points with an intensity gradient method and described feature points by 3D HoG. The intensity gradient (IG) method relaxes the severe constraints of a 3D Harris detector. Instead of computing the second moment matrix, IG computes the sum of all three gradients in spatial and temporal dimensions. This results in interest points along edges, which are moving. This method is simple but produces a good number of interest points. Prior experimental result showed that this approach to feature point extraction can achieve a result comparable to 3D Harris corners and temporal Gabor filter [17]. As a descriptor, 3D HoG is extracted to represent space-time cuboids. The main differences between IG+3D HoG and MoSIFT are the scale invariance and the more comprehensive descriptor in MoSIFT.

Our method does not require any additional information such as manual "hot spot" masking, people detection, or people tracking. MoSIFT also does not impose limitations as to how many actions can happen at once or require detailed annotations that mark where an action takes place. Table 3 shows that MoSIFT yields promising results in recognizing human actions in real world surveillance videos with an average relative increase of 22.5% to an absolute accuracy of 38% compared to prior work [18]. MoSIFT improves recognition of all actions except for CellToEar, where motion of moving the hand with a cell phone to the ear is exceedingly difficult to spot, even for a person. Several actions with clearly visible motions show dramatic increases in absolute recognition accuracy such as putting an object down (ObjectPut to 43%), people meeting (PeopleMeet to 45%), people splitting up (PeopleSplitUp to 55%) and running (PersonRuns to 36%). Sample actions and MoSIFT detected points are shown in Figure 6.

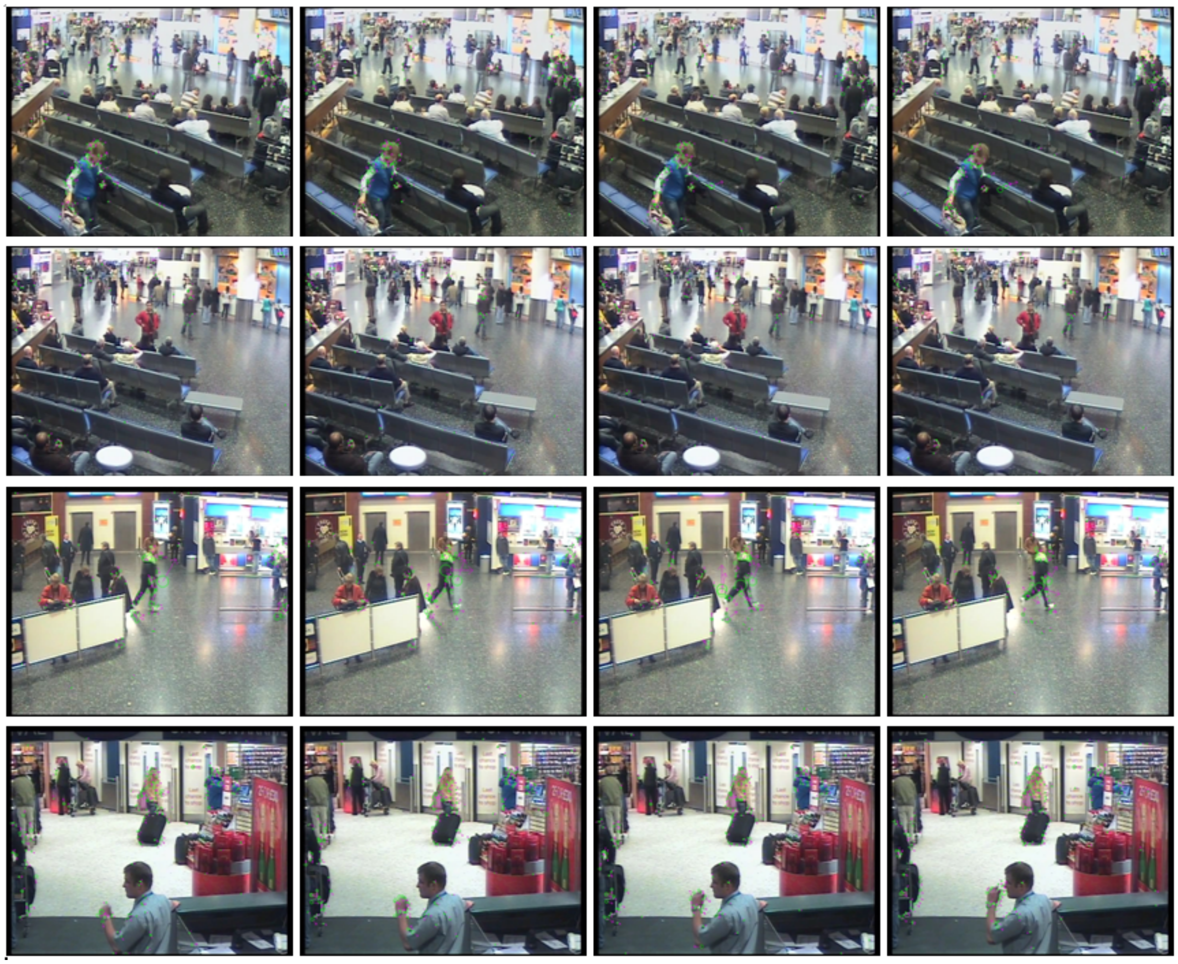


Figure 6: Some examples of MoSIFT detection. Top row shows a ObjectPut action. The second row is a CellToEar action (by the person sitting in the bottom left corner), the 3rd row is a PersonRuns action and the bottom row is PeopleMeet. Green circles indicate MoSIFT interest points and purple arrows show the direction in which the interest points are moving.

We further combined top 500 bigrams from tf-idf ranking with MoSIFT. The bigram model slightly improved MoSIFT algorithm with additional global information.

5. Conclusion and future works

The paper has presented a new method, MoSIFT, to detect interest points and describe local features for human action recognition. MoSIFT treats the spatial domain and the temporal domain separately. Interest point detection is based on spatial appearance and 'sufficient' motion. The feature descriptor captures both local appearance and motion as histograms of gradients in space and histograms of optical flow. The MoSIFT descriptor is made more robust through grid aggregation of both histograms. Pair-wise correlation is encoded as a bigram model with MoSIFT to provide additional structure and sequence information. This combination of techniques improves the state of art results on KTH to an accuracy of 95.8%. We also demonstrated that MoSIFT shows good results in a large and complex real world surveillance dataset.

The most challenging next step for us is to extend our method to action detection beyond the already respectable action recognition (classification). Additionally, we would like to not only detect behaviors but also locate them in the frame, something for which the bag of words approach is not well suited.

Reference

[1] I. Laptev, and T. Lindeberg. Space-time interest points, In ICCV, p. 432-439, 2003

[2] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In ICPR, 32-36, 2004

[3] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In ICCV, 166-173, 2005

[4] P. Dollár, V. Rabaud, G. Gottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features, In VS-PETS 2005, page 65-72

[5] J. C. Nibles, H. Wang, and L. F.-F. Li. Unsupervised learning of human action categories using spatial-temporal words. In BMVC, 2006.

[6] S.-F. Wong, and R. Cipolla. Extracting spatiotemporal interest points using global information. In ICCV 2007.

[7] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. IEEE Trans. Systems, Man, and Cybernetics, Part B, 36(3):710-719, June 2006.

[8] National Institute of Standards and Technology (NIST): TRECVID 2008 Evaluation for Surveillance Event Detection. http://www.nist.gov/speech/tests/trecvid/2008/doc/EventDet08-EvalPlan-v04.htm, 2008

[9] D.G. Lowe. Distinctive image features from scale invariant key points, In IJCV, November 2004

[10] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In CVPR 2008

[11] K. Schindler, and L.V. Gool. Action Snippets: How many frames does human action recognition require? In CVPR 2008

[12] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In IJCV, 2007

[13] R. Bekkerman, and J. Allan. Using Bigrams in Text Categorization. CIIR Technical Report IR-408 2004

[14] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlations. In CVPR 2006

[15] S. Savarese, A.D. Pozo, J.C. Niebles and F.-F. Li. Spatial-temporal correlations for unsupervised action classification. IEEE Workshop on Motion and Video Computing. , 2008

[16] Agarwal, S., Awan, A., and Roth, D. 2004, Learning to detect objects in images via a sparse, part-based representation, PAMI, November 2004

[17] C. Chen, H. Wactlar, M.-Y. Chen, G. Can, A. Bharucha, and A. Hauptmann. Recognition of Aggressive Human Behavior Using Binary Local Motion Descriptors. International IEEE Engineering in Medicine and Biology Conference (EMBC), 2008

[18] A. Hauptmann, R. V. Baron, M. Chen, M. Christel, W.-H. Lin, X. Sun, V. Valdes, J. Yang, L. Mummert, and S. Schlosser Informedia @ TRECVID2008: Exploring New Frontiers TRECVID Video Retrieval Evaluation Workshop, NIST, Gaitherburg, MD, November 2008