Dynamic Vision Sensors for Human Activity Recognition

Abstract

Unlike conventional cameras which capture video at a fixed frame rate, Dynamic Vision Sensors (DVS) record only changes in pixel intensity values. The output of DVS is simply a stream of discrete ON/OFF events based on the polarity of change in its pixel values. DVS has many attractive features such as low power consumption, high temporal resolution, high dynamic range and less storage requirements. All these make DVS a very promising camera for potential applications in wearable platforms where power consumption is a major concern. In this paper we explore the feasibility of using DVS for Human Activity Recognition (HAR). For HAR we propose to use the various slices (such as $x-y$, $x-t$ and $y-t$) of the DVS video as a feature map and denote them as Motion Maps. We show that fusing motion maps with Motion Boundary Histogram (MBH) gives good performance on the benchmark DVS dataset as well as on a real DVS gesture dataset collected by us. Interestingly, the performance of DVS is comparable to that of conventional videos although DVS captures only sparse motion information.

1. Introduction

Conventional video camera uses frame based visual acquisition where each pixel is sampled at a fixed frame rate irrespective of whether or not their value changed. This leads to a lot of data redundancy and hence increased bandwidth and memory requirements. Dynamic Vision Sensor (DVS) [9] is a recent innovation in machine vision that mimics some of the functionalities of the human retinal vision. Instead of capturing the whole frame, it records only those pixels that see a change in intensity values. If the magnitude of change in log intensity value at a pixel is above a particular threshold, then it generates either an ON or an OFF event.

A major advantage of DVS is its ultra low power consumption. This is because it only generates ON/OFF events and hence avoid the use of ADCs which consumes the most power in conventional cameras. Hence DVS could be used to boost the battery life in wearable or portable devices like untethered Augmented Reality (AR) devices, which currently use conventional cameras for various purposes such as gesture/activity recognition and building $3-\mathrm{D}$ maps. With this idea in mind, we explore performing activity/gesture recognition using DVS.

Another reason why DVS is intrinsically suitable for gesture/activity recognition is because it does not record any static information about the scene. Thus, we can avoid the overhead of many preprocessing algorithms used in conventional image processing such as background subtraction and contour extraction.

For the task of human activity recognition, we propose a simple method of using various slices ($x-y$, $x-t$ and $y-t$) of the DVS video as feature maps. We denote these maps as *motion maps* and employ Bag of Visual Words framework to extract critical features. Recognition rates obtained were similar to that of existing descriptors under this setting. We also combined the motion maps' features with state-of-the-art motion descriptor Motion Boundary Histogram (MBH) to obtain the best recognition rates, much higher than the HAR performance of individual descriptors. The results on DVS data are even comparable with the recognition rates seen in conventional videos. This is quite surprising given that DVS data is a very compressed version of the original video data with a remarkably sparse encoding. We experimented on two datasets: the DVS recordings of UCF11 [7] and a hand gesture DVS dataset collected by us. In both the datasets our results have been promising for DVS.

1.1. Related Work

There are several works in literature for human activity recognition, of which we mention here a few relevant ones. For a typical activity recognition task, two types of features are classically extracted - descriptors based on motion and those based on shape. Motion History Images (MHI) from videos accumulates foreground regions of a person and accounts for its shape and stance [3]. Several more contour-based approaches such as Cartesian Coordinate Features (CCF), Fourier Descriptors Features (FDF) [8, 6], Centroid-Distance Features (CDF), and Chord-Length Features (CLF) provides shape description [19]. For motion based descriptors, Histogram of Optical Flow (HOF) computes optical flow of pixels between consecutive frames using brightness constancy assumption [11, 4]. Motion boundary histograms take one step further by performing derivative operation on the optical flow field. This makes the feature invariant to local translation motion of the camera and captures only relative motion in the video [5, 17].

Several other descriptors work by extracting the scene (background), color/hue and texture based features in a video. But texture and hue information is unavailable in DVS data because of its binary encoding scheme. The scene context based descriptors can also not be used with DVS videos since scenes usually are static in a video, unless there is significant camera motion. Nevertheless, volume based features like motion and shape often provide sufficient information required to perform decent recognition and are more popular than surface features like color and texture.

Human activity recognition has been popularly solved by extracting local features from videos on which Bag of Visual Words model (BoVW) is learnt and a classifier, typically SVM is trained [18, 14]. [14] describes different good practices for extracting BoVW and several fusing techniques to combine descriptors in order to produce state-of-the-art recognition system. As against these simple methods, recent works on HAR has focussed on deep learning techniques for improving recognition rates. Deep Convolutional and LSTM Recurrent Neural network units can be trained to automate feature extraction and directly perform natural sensor fusion [13] on human videos. Two-stream Convolutional Neural Networks learn the spatial and temporal information extracted from RGB and optical flow images of videos and are also becoming common for activity recognition [12, 15].

Since DVS recording provide us with both motion and shape cues, in this paper, we exploit these critical information by proposing a fusion of simple shape and motion based feature descriptors.

2. DVS Based Activity Recognition

Unlike conventional camera, DVS captures only motion and thus avoids the need to perform background subtraction. Also DVS output is usually sparse because change in pixel intensity occurs only at texture edges. To exploit this sparsity and motion cues captured by DVS, we propose to extract various projections of the DVS data (the motion maps) and use them as feature descriptor for activity recognition. Finally we fuse motion maps and with a state-of-the-art motion descriptor MBH [5] to further improve the recognition accuracy. The overall architecture is shown in Figure 1.

We first convert DVS event streams into a video by accumulating events over a time window. For example, for our experiments we converted the DVS data into a $30fps$ video. From this video, we obtain three different 2-D projections: $x - y$, $x - t$ and $y - t$ by averaging over the left-out dimension in each case. Thus, $x - y$ projection is obtained by averaging over the $time$ axis, $x - t$ by averaging over the $y - axis$ and $y - t$ by averaging over the $x - axis$. We call these 2-D projections as motion maps since DVS captures the direction of motion of the foreground object. The $x - y$ motion map gives us the mean pose and stance of an object in the scene whereas the $x - t$ and the $y - t$ motion maps record the manner in which the object had moved over the video's duration. Our proposed $x - y$ motion map is similar to the idea of motion history images [1] but we have two additional maps that account for the movement of the object along the horizontal and vertical directions.

From the three motion maps, we extract Bag of Features (BoF), where we use Speeded Up Robust Features (SURF) [2] extracted through grid search on the maps. This is followed by $k - means$ clustering of the training data's features to create a visual vocabulary of k words. Then the features from each video are binned to these k clusters and are L_2 normalized. Finally, a *linear* SVM classifier under *one-vs-all* encoding scheme is trained on the encoded features to predict the performed activity. Since the three motion maps inherently complement each other with the $x-y$ map encoding the shape and pose of the object while the $x - t$ and the $y-t$ motion maps describing its motion, we combine all the three motion maps' descriptors to obtain better classification accuracy. We tried fusion of features before as well as after performing BoF and observed that fusion after BoF performs better. This result is in line with that of [14] where the authors have given a comprehensive study of fusing feature descriptors at several stages of BoF. The authors in [14] also concluded that concatenating feature encoding of separately trained BoF models before training a single classifier gave the best recognition results and named this fusion as representation level learning.

For the final recognition task, the motion maps descriptors are further combined with the MBH descriptor since MBH also encodes the appearance of objects and local motion in video but in a method that is distinctly different. MBH takes derivative of optical flow which in turn is computed using derivative of the video frames with respect to its spatial and temporal coordinates. Hence MBH employs second order statistics for its feature extraction while the motion maps use simple zero order statistics. Thus these two descriptors supplement one another and their combined feature set outperforms the individual recognition rates. In the next section, we evaluate these descriptors and the loss in performance of HAR on using DVS data compared to conventional videos on a benchmark dataset. We also report the performance on the DVS gesture dataset we have collected. From the results, we assess the usability of DVS for activity recognition and conclude with its shortcomings.

3. Datasets

This section describes the datasets we used for testing and evaluating our descriptors. Experiments were carried

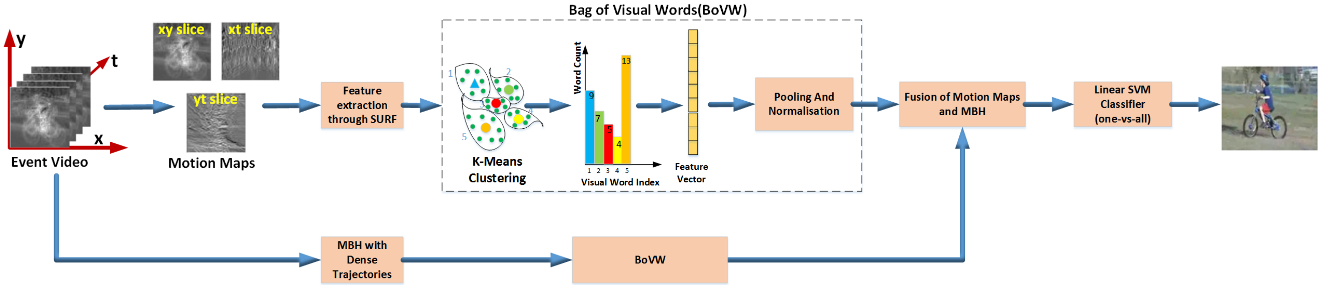

Bag of Visual Words(BoVW)
y
t
x
Event Video
xy slice
xt slice
yt slice
Motion Maps
Feature extraction through SURF
K-Means Clustering
Word Count
13
9
7
4
1 2 3 4 5
Visual Word Index
Feature Vector
Pooling And Normalisation
Fusion of Motion Maps and MBH
Linear SVM Classifier (one-vs-all)
MBH with Dense Trajectories
BoVW

Figure 1: Our proposed method: The event stream from DVS is converted to a video at $30fps$. Motion maps are generated through various projections of this event video and SURF features are extracted from these motion maps. MBH features using dense trajectory are also extracted from the event video. Bag of features encoding from both these descriptors are combined and given as input to linear SVM classifier (*one-vs-all*).

out on two datasets, namely the UCF YouTube Action Data Set or UCF11 [10] and a DVS gesture dataset collected by us using DVS128.

The UCF11 data was chosen because it is one of the few human action datasets whose benchmark DVS counterpart is publicly available [7]. The DVS data was created by the authors [7] by re-recording the existing benchmark UCF11 videos played on a monitor using a DAViS240C vision sensor. Since the data was not directly recorded from the wild, this would mean that time resolution greater than that provided by the UCF11 video is not available in DVS under this simulated setting. Nonetheless, the dataset is sufficient for our experiments since it captures the sensor noise in DVS and is used on action videos that by themselves are not very fast paced.

The UCF11 dataset contains eleven action classes as shown in Figure 2, viz. basketball shooting, biking, diving, golf swinging, horse riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking and walking dog. It has more than 100 videos in each class which is further subdivided in to 25 groups. This grouping allows us to perform Leave One Out (LOO) cross validation twenty five times on the actions as suggested by the creators of the data.

Figure 2: YouTube Action Data Set [1]

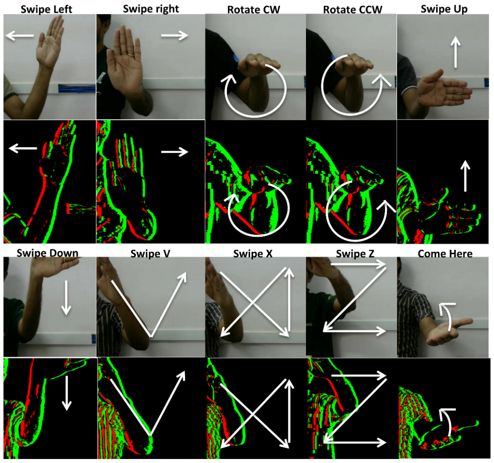


Figure 3: Gestures from the DVS dataset collected by us. Ground truth from an RGB camera is also shown for reference.

For a second round of experiments, we used the DVS hand gesture dataset collected by us using DVS128 and we refer them as the *DVS Gesture data*, see Figure 3. The dataset contains 10 different hand gestures, each performed 10 times by 12 subjects constituting a total of 1200 gestures. The hand gestures used are left swipe, right swipe, beckon, counter-clock wise rotation, clock wise rotation, swipe down, swipe up, swipe V, wave X and wave Z. We performed 12-fold cross-validation for all experiments on this dataset leaving out one subject each time for testing. Links to the dataset as well as the code for our proposed method will be made available in the final paper.

For an illustration of the motion maps, see Figure 4 where we have created the maps from randomly picked videos of the eleven classes of UCF11 data. Note that in the $x - y$ map, much of the shape and pose of the object is

[1] Image source: http://crcv.ucf.edu/data/UCF_YouTube_Action.php

captured. Similarly, the $x - t$ and $y - t$ slices show rhythmic patterns based on the movement involved typical for a given action category. Notable ones among these are winding river-like $y - t$ motion map for action class *swinging* and the rhythmic up and down spikes in the $x - t$ motion map for *trampoline* class.

4. Feature Extraction and classification

This section describes the steps that we used for feature extraction and classification. To evaluate existing motion descriptors like HoG, HOF and MBH, we extracted local spatio-temporal features using dense trajectories from the videos [16]. Dense trajectories are created by dense sampling of images frame wise. The sampled points are tracked along frames using dense optical flow field and the trajectory length is limited to 15 frames to avoid drifting of tracked points. Along the path tracked, descriptors like HoG, HOF or MBH are computed using the neighborhood volume of $32 \times 32 \times 15$ pixels. This volume is further divided into cells of size 16×16 pixels $\times 5$ frames. So each tracked tube gives a $2 \times 2 \times 3$ cells. Within each cell, the histograms of descriptors are found. For HoG and MBH, we used 8 orientation bins per cell and the magnitude of the feature values were used for weighting. For HOF, an additional bin was added to account for pixels whose flow value was smaller than a threshold. All the descriptors were also L_2 normalized before performing bag of features. In total, HoG gave feature descriptors of size 96 per tracked volume ($2 \times 2 \times 3$ cells per tracked path times 8 bins) while HOF produced 108 features ($2 \times 2 \times 3$ cells times 9 bins). MBH also gave 96 features similar to HoG, but in both horizontal and vertical directions. Thus, overall it had twice the number of features for a chosen trajectory. Bag of features was individually performed on each of these descriptors. Since each video produced about $\approx 500,000$ dense tracks, most of them in close proximity to one another, BoF was done on a subset of training features on $100,000$ trajectories randomly selected. To ensure that every video in the train set contributes to the codebook, we have selected a fixed number of features randomly from each video instead of first pooling all the extracted features and then performing random selection. The codebook dimension in the clustering step was maintained at 500. After learning the cluster centers, all features of the video were used to generate the histograms of the same 500 bins. Finally the segregated features were L_2 normalized and SVM classifier was trained.

On the motion maps also, we performed bag of features individually with a codebook of dimension 500. We have used Matlab's built-in function `bagOfFeatures` for this step, followed by training *one-vs-all linear* SVM for the multi-class recognition. The results under Leave One Out cross-validation method for all these descriptors are given in the next section.

5. Experimental Results

We have conducted our experiments on two datasets as explained in the following section.

5.1. HAR on UCF11 and its DVS counterpart

In this experiment, HAR was performed on the original UCF11 dataset (RGB) and its corresponding DVS recordings. Table 1a provides the recognition rates obtained with 25 fold Leave One Out cross-validation method.

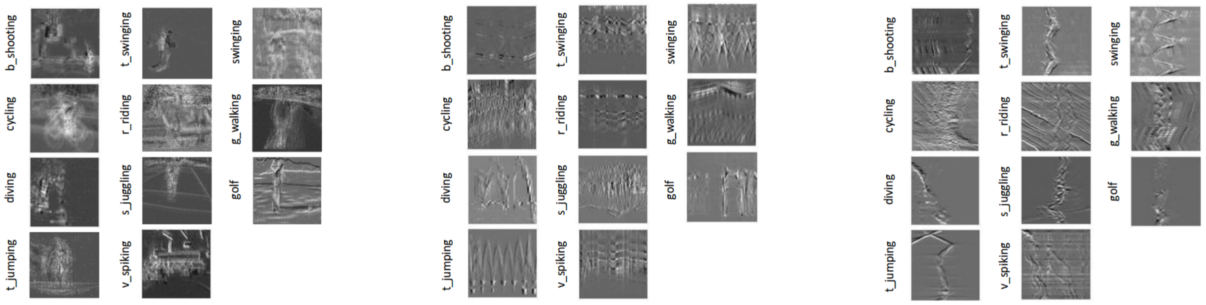
The results show that fusion of motion maps from the DVS data gave a HAR rate of 67.27%, which is comparable to the rates for HOF and HoG on the original UCF11 data. Interestingly, with the individual motion map descriptors the DVS recordings of UCF11 gave higher recognition rates while descriptors like HoG and HOF performed better on the original videos. The reason why the motion maps worked better on the DVS data is that there is no background clutter and scene information in DVS recording for distracting its bag of features encoding. KNN ($k-$nearest neighbors) classifier was also used for the final predictions, but it gave consistently about 5% lower HAR rates. Similarly, it was also observed that simply using a larger dimension codebook of size 4000 improves recognition rates by $2 - 3\%$. Because our aim is to study the performance of DVS data for HAR compared to original data, we limited our codebook size to 500 words for all our experiments as using higher sized codebook simply improved both the results.

In order to further boost HAR rates, we separately included the MBH as well as HOF descriptors along with the motion maps in light of the complementarity it offers. SVM classifier under *one-vs-all* linear encoding was trained to get the best recognition rates. The HAR values in Table 1a show that the features from motion maps better complement the second order statistics of MBH than the first order HOF features on both the UCF11 datasets, as evidenced by the greater accuracy of combined MBH features and motion maps. The results also show that this final fusion gave the highest recognition rate among all descriptors and nearly bridged the performance gap between DVS and conventional videos on the benchmark UCF11 data. Given the sparsity of DVS data, it is remarkable that the final descriptor has provided a near-equivalent performance on DVS when compared to the results on conventional videos.

5.2. Recognition on our DVS gesture dataset

In the second round, we analyzed the recognition rates on the DVS gesture dataset collected by us.

With our DVS gesture dataset also, we obtained decent recognition rates by combining the motion maps alone, nearly same as that given by existing feature descriptors. Combining motion maps with the MBH descriptor again



(a) $x - y$ Motion Map

(b) $x - t$ Motion Map

(c) $y - t$ Motion Map

Figure 4: The three Motion maps for 11 randomly picked UCF11 videos. Figures (a), (b) and (c) show their motion maps averaged along t, y and x axes respectively.

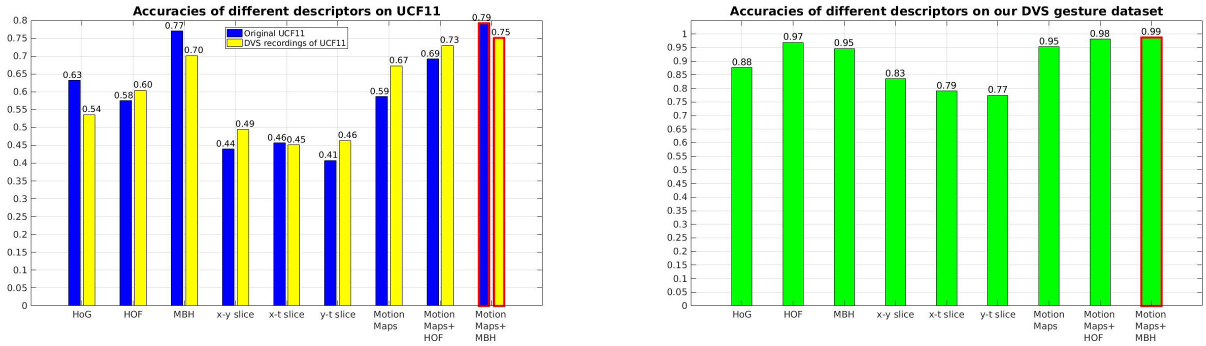
Dataset	HoG	HOF	MBH	*x-y* Motion Map	*x-t* Motion Map	*y-t* Motion Map	Combined Motion Maps	Motion Maps + HOF	Motion Maps + MBH
Original UCF11	0.6319	0.5754	0.7707	0.4397	0.4567	0.4077	0.5867	0.6922	**0.7933**
DVS recordings of UCF11	0.5358	0.6043	0.7016	0.4943	0.451	0.4629	0.6727	0.7299	**0.7513**

(a) Results on UCF11 and its DVS counterpart

Dataset	HoG	HOF	MBH	*x-y* Motion Map	*x-t* Motion Map	*y-t* Motion Map	Combined Motion Maps	Motion Maps + HOF	Motion Maps + MBH
DVS gesture dataset	0.8768	0.9689	0.9468	0.7748	0.8349	0.7899	0.9529	0.9809	**0.9880**

(b) Results on the DVS gesture dataset collected by us

Table 1: Recognition rates for various motion and shape descriptors on the UCF11 dataset, its corresponding DVS data and on the DVS gesture dataset collected by us.



(a) Accuracies on original UCF11 and its DVS counterpart

(b) Accuracies on our DVS gesture dataset

Figure 5: Accuracy plots on UCF11 and our DVS gesture dataset.

augmented HAR performance to give the highest recognition rates as seen in Table 1b.

6. Conclusion and Future Work

In this paper, we have analyzed the performance of DVS data in human activity recognition and compared it with its conventional frame-based counterpart using traditional feature extraction techniques. We also proposed a new encoding technique (motion maps) that is suited especially for DVS data in light of its sparse and concise recording scheme. We noted how combining the existing MBH descriptor with motion maps gave the best recognition results (Figure 5). Our results have shown that DVS for HAR is a promising application.

Based on the feature descriptors available for its encoding, HAR results from DVS recordings have been nearly equal to that of RGB videos on the benchmark UCF11 data. Additional features based on the scene, texture and hue have enabled better recognition rates with actual videos. But these are more complex and unavailable for use with the DVS data from the very beginning. Hence respecting the limitations that come with DVS, we conclude that within the framework of its possible descriptors it is just as useful for HAR as conventional videos and could efficiently be used in place of the latter, especially in low power and high speed applications.

As future work, we can look at improving performance of simple bag-of-features where location based relations are not preserved due to its pooling step. Rather than destroying spatial information between the extracted features in the image, methods like *Spatial Correlogram* and matching can be employed on the DVS data. Also, we noted that similar to recognition rates shown by conventional videos, dense trajectories with MBH gave the best results on using traditional features in DVS as well. Much of the success of dense tracking comes from the fact that it generates too many interest points given any video sample. Visualization of the interest points found by dense sampling showed that some of these are randomly fired noisy events in DVS unrelated to the object in foreground. A simple median filtering preprocessing before finding dense interest points however *did not* improve recognition rate. In order to truly address the problem, a new method specifically for finding and tracking DVS events should itself be invented. This would act as the true initial step for improving the performance of HAR on using optical flow, MBH as well as dense trajectories with dynamic vision sensors.

References

[1] M. A. Ahad. Motion history image. In *Motion history images for action recognition and understanding*, pages 31–76. Springer, 2013. 2

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer vision–ECCV 2006*, pages 404–417, 2006. 2

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001. 1

[4] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International journal of computer vision*, 61(3):211–231, 2005. 1

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, pages 428–441. Springer, 2006. 2

[6] R. D. De Leon and L. E. Sucar. Human silhouette recognition with fourier descriptors. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 3, pages 709–712. IEEE, 2000. 1

[7] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in Neuroscience*, 10, 2016. 1, 3

[8] H. Kauppinen, T. Seppanen, and M. Pietikainen. An experimental comparison of autoregressive and fourier-based descriptors in 2d shape classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):201–207, 1995. 1

[9] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128x128 120 db 15 μ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1

[10] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, pages 1996–2003. IEEE, 2009. 3

[11] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981. 1

[12] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *arXiv preprint arXiv:1703.10667*, 2017. 2

[13] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. 2

[14] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016. 2

[15] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2

[16] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011. 4

[17] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013. 2

[18] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007. 2

[19] D. Zhang and G. Lu. Review of shape representation and description techniques. *Pattern recognition*, 37(1):1–19, 2004. 1