A 3-Dimensional SIFT Descriptor and its Application to Action Recognition

Paul Scovanner
Computer Vision Lab
University of Central Florida
pscovanner@cs.ucf.edu

Saad Ali
Computer Vision Lab
University of Central Florida
sali@cs.ucf.edu

Mubarak Shah
Computer Vision Lab
University of Central Florida
shah@cs.ucf.edu

ABSTRACT

In this paper we introduce a 3-dimensional (3D) SIFT descriptor for video or 3D imagery such as MRI data. We also show how this new descriptor is able to better represent the 3D nature of video data in the application of action recognition. This paper will show how 3D SIFT is able to outperform previously used description methods in an elegant and efficient manner. We use a bag of words approach to represent videos, and present a method to discover relationships between spatio-temporal words in order to better describe the video data.

1. INTRODUCTION

Action recognition is a well studied yet very difficult problem in the task of automatically understanding video data. Intra-class variation is often very large and confusion is common between actions such as running and jogging. Actions depicted by video data inherently contain spatio-temporal information, which implies that descriptors are needed which can robustly encode this kind of information. A few example actions are shown in Figure 1. In the past, solutions to the action recognition problem have utilized features such as optical flow [4], silhouette shapes [12], volume based representation [5], etc. In the past few years, the bag of words paradigm [3] has shown remarkable performance for image classification and object detection in single images. Originally this paradigm was inspired by the bag of words approach to text categorization. The key reason that contributed towards the success of this form of processing for image classification was the usage of robust SIFT [8] descriptors for representing image regions. In this paper we extend the bag of words paradigm from 2 dimensions (2D) to 3D, where the third dimension is time, and demonstrate its application for the task of action recognition.

Previous methods which extend bag of words to video have tested only simple features such as gradient magnitude [10]. However, note that these features do not explicitly describe the true spatio-temporal nature of the video data. The 3D SIFT descriptor encodes the information local in both space and time in a manner which allows for robustness to orientations and noise. In addition, after describing the videos as a bag of spatio-temporal words using the proposed SIFT descriptor, we discover relationships between words to form spatio-temporal word groupings. A co-occurrence based criteria is used for this purpose. The discovered groupings are finally used for the classification task. Finally in the experiments section of this paper, we demonstrate the superior performance of our proposed framework based on 3D SIFT descriptor for the task of action recognition.

Before presenting the details of our method we would like to summarize the novel contribution of our paper. The contributions are:

- Formulation of 3D SIFT descriptor that accurately captures the spatio-temporal nature of the video data.
- Extension of bag of words paradigm to videos using a framework based on 3D SIFT.
- An algorithm to discover relationships between spatio-temporal words for learning discriminative word groupings.
- Comparative analysis of our 3D SIFT descriptor with previous descriptors used for the same purpose.

2. PREVIOUS WORK

Much work has been done using the original SIFT descriptor [8, 9] for tasks such as object recognition [8, 9], point tracking [11, 7], panorama creation [2], etc. This work has concentrated on matching interest points between static images, a task in which SIFT excels. Previous work has

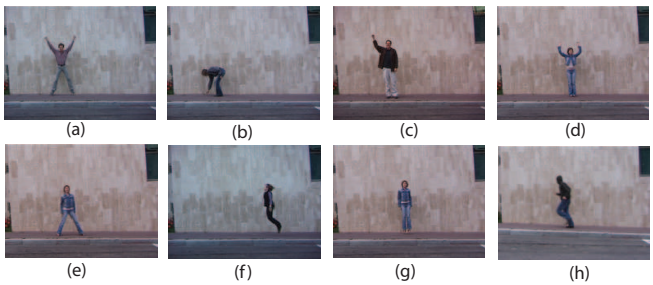

Figure 1: Example Actions from the test data set (a) Jumping-jack (b) Bend (c) Single handed wave (d) Two handed wave (e) Side-skip (f) Jump (g) Stationary jump (h) Run.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

also focused on the use of bag of words for action recognition in videos. For instance, in [10] videos of actions are classified using a bag of words approach. In this method, interest points are detected by searching the entire video for local maxima to a Gabor response function at a given scale. Features used to describe these points are simply the vectorized gradient values of the 3D neighborhood surrounding the point. In our method, interest points are chosen at random, allowing for faster runtime and the ability to be used in an online manner. Also, a new feature descriptor is used. This 3D SIFT descriptor is able to robustly describe the 3D nature of the data in a way that vectorization of a 3D volume can not. Using sub-histograms to encode local time and space information allows 3D SIFT to better generalize the spatio-temporal information than features used in previous works. We will show that the 3D SIFT descriptor outperforms the descriptors used by previous methods of action recognition. All of this translates into a fast and accurate method of action recognition.

3. 3D SIFT DESCRIPTOR

This section will outline the differences between the 2D SIFT descriptor and the 3D SIFT descriptor and discuss their impact. The first step is to compute the overall orientation of the neighborhood. Once this is computed we can create the sub-histograms which will encode our 3D SIFT descriptor.

3.1 Orientation Assignment

The 2D gradient magnitude and orientation for each pixel is defined as follows:

$$m_{2D}(x,y) = \sqrt{L_x^2 + L_y^2}, \quad \theta(x,y) = \tan^{-1}(\frac{L_y}{L_x}). \quad (1)$$

where L_x and L_y are respectively computed using finite difference approximations: $L(x+1,y,t) - L(x-1,y,t)$ and $L(x,y+1,t) - L(x,y-1,t)$. Similarly, in 3D (x, y and t), the spatio-temporal gradient $(L_x, L-y, L-t)$ can be computed, where L_t is approximated by $L(x,y,t+1) - L(x,y,t-1)$. Now the gradient magnitude and orientations in 3D are given:

$$m_{3D}(x,y,t) = \sqrt{L_x^2 + L_y^2 + L_t^2}, \quad (2)$$

$$\theta(x,y,t) = \tan^{-1}(L_y/L_x), \quad (3)$$

$$\phi(x,y,t) = \tan^{-1}(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}). \quad (4)$$

It can be observed that ϕ now encodes the angle away from the 2D gradient direction. Due to the fact that $\sqrt{L_x^2 + L_y^2}$ is positive, ϕ will always be in the range $(-\frac{\pi}{2}, \frac{\pi}{2})$. This is a desired effect, causing every angle to be represented by a single unique (θ, ϕ) pair. Each pixel now has two values which represent the direction of the gradient in three dimensions. The next step is to construct a weighted histogram similar to that of [8] for the 3D neighborhood around a given interest point. There are multiple ways of accomplishing this. One way is by dividing θ and ϕ into equally sized bins (creating meridians and parallels) and creating a 2D histogram, another is to tessellate the sphere using a icosahedron. In this paper we use the meridians and parallels method since it is much simpler and faster to find peaks of a 2D orientation histogram, and quadratically interpolate the true peaks.

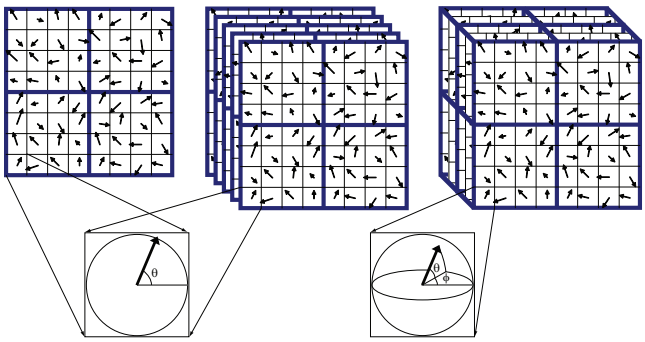

Figure 2: The left image shows the familiar 2D SIFT descriptor. The center shows how multiple 2D SIFT descriptors could be used on a video without modification to the original method. The right shows the 3D SIFT descriptor with its 3D sub-volumes, each sub-volume is accumulated into its own sub-histogram. These histograms are what makes up the final descriptor.

One point to note is that when using the meridians and parallels method, bins will need to be normalized by their solid angle (ω). This is required to correct for a problem that is apparent to anyone who has looked at a map of the world. Any 2D map of the earth must either stretch areas near the poles, or create discontinuities. We must normalize the values added to each bin by the area of the bin, also called the solid angle. If one were to skip this step, the orientation histogram would be incorrectly weighted towards the equator. The solid angle can be calculated in the following manner:

$$\omega = \int_{\phi}^{\phi+\Delta\phi} \int_{\theta}^{\theta+\Delta\theta} \sin\theta \; d\theta \; d\phi \quad = \Delta\phi \int_{\theta}^{\theta+\Delta\theta} \sin\theta \; d\theta$$
$$= \Delta\phi \left[-\cos\theta\right]_{\theta}^{\theta+\Delta\theta} \quad = \Delta\phi \left(\cos\theta - \cos(\theta+\Delta\theta)\right).$$

The actual value added to the histogram is shown below, where (x,y,t) represents the location of the interest point, and (x',y',t') represents the location of the pixel being added to the orientation histogram. The peaks of this histogram therefore represent the dominant orientations. The dominant peak is stored as it can be used to rotate the neighborhood around the key point, creating rotationally invariant features.

$$hist(i_\theta, i_\phi) += \frac{1}{\omega} m_{3D}(x',y',t') e^{\frac{-((x-x')^2+(y-y')^2+(t-t')^2)}{2\sigma^2}} \quad (5)$$

3.2 Descriptor Representation

The next step is to compute the SIFT descriptor for which we start by calculating the orientation sub-histograms. The first step in this process will be to rotate the 3D neighborhood surrounding the key point so that the dominant orientation (calculated in the Orientation Assignment stage) points in the direction of $\theta = \phi = 0$. This is done by taking each (x,y,z) position in the neighborhood and multiplying it by the following matrix

$$\begin{bmatrix} \cos\theta\cos\phi & -\sin\theta & -\cos\theta\sin\phi \\ \sin\theta\cos\phi & \cos\theta & -\sin\theta\sin\phi \\ \sin\phi & 0 & \cos\phi \end{bmatrix}. \quad (6)$$

To create our sub-histograms we sample 8 sub-regions surrounding the interest point as shown in Figure 2 (4x4x4 pixel

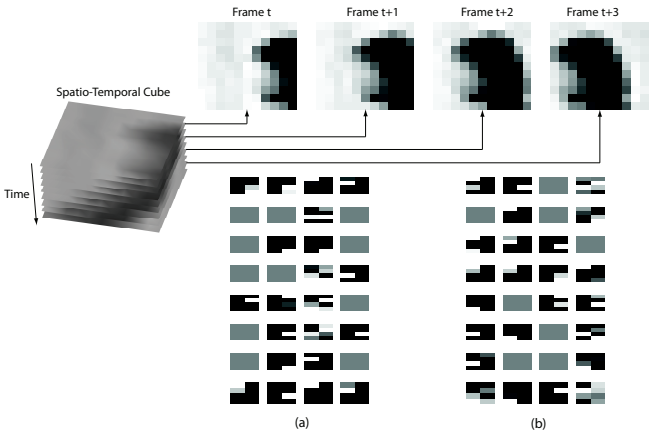

Figure 3: Visualization of the abbreviated descriptor used to view the feature vector itself. Each of the 8x4 sub-plots represent an orientation bin, and each gray value in these subplots represents the value of the 2x2x2 sub-histogram (reshaped to 4x2). (a) Shows the descriptor before global reorientation by the overall maximum orientation direction. (b) Shows the descriptor after global reorientation.

regions are used in experimentation but for visual aesthetics fewer are shown in the figure), where each pixel contains a single magnitude value and two orientation values θ and ϕ. What was originally trilinear interpolation, now becomes quintalinear (five dimensional) interpolation.

4. ACTION CLASSIFICATION

In this section we will describe the steps involved in our proposed action classification framework. The first step is to select the salient regions from the spatio-temporal video cube. For this purpose we carry out random sampling of a video at different locations, times, and scales. Note that interest points could also be extracted from video content using other methods [10]. However, these methods require additional processing stages which can be costly. Once the points are sampled the second step is to describe the spatio-temporal region around the points using the proposed 3D SIFT descriptor. The length of the descriptor is based on the number of sub-histograms, and the number of bins used to break represent the θ and ϕ angles. In our case we used **2x2x2** and **4x4x4** configurations of sub-histograms, and **8x4** histograms to represent θ and ϕ. This yields descriptors of length **256** and **2048** dimensions. We observed slight improvements when using the larger feature vectors and the results of this paper use the larger descriptor, however the abbreviated feature descriptor could be used to improve runtime test speeds. The descriptors gathered from all the interest points are then quantized by clustering them into a pre-specified number of clusters. Figure 4 pictorially describes this process. This step is carried out in an unsupervised manner using a hierarchical k-means clustering method. The resultant cluster centers are now called 'words', while the collection of these cluster centers is referred to as the 'spatio-temporal word vocabulary'.

Now that our vocabulary is computed, the 3D SIFT descriptors from the videos are matched to each 'word' and the frequency of the words in each video is accumulated into a

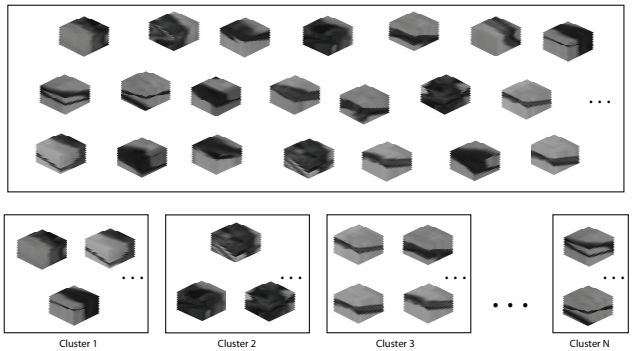

Figure 4: Visual representations of the words. The top box contains cubes randomly sampled from the data set. The clusters below show how similar video cubes are grouped together based on the similarity of their feature representations.

histogram. This word frequency histogram, referred to as a 'signature', is used to generate an initial representation of the video. We do not use these signature directly, instead we create a feature grouping histogram out of it. The reason for this step is the observation that for any particular action some words co-occur more then others. If we can discover such co-occurrences among the words, they can be used to build a more discriminative representation of the given action video.

The test for finding word co-occurrences is carried out as follows. We construct a word co-occurrence matrix and populate it using frequency histograms of videos. If the size of the vocabulary is N then the co-occurrence matrix will have dimensions N x N. Each row vector in this matrix stores how many times a particular word occurred with any other word. This row vector can also be considered as representing the contextual distribution of that word in terms of other words of the vocabulary. Now the observation is that if any two words have similar contextual distributions for a particular action, that means these two words are capturing something similar and therefore are related to each other. For quantifying this observation we compute correlation between the distribution vectors of any two words. If the correlation is above a particular threshold we join them together and add their corresponding frequency counts from their initial histograms into a new grouping histogram. In this new grouping histogram each bin corresponds to one such grouping. Finally, SVM learning is used to train representative models for each action category using grouping histograms as feature vectors. Separate SVM classifiers are trained for each action, the testing video is classified by the classifier which has the greatest distance from the SVM hyperplane.

5. EXPERIMENTS

The task of our method is to classify actions present in the given video. For this purpose we employ the action data set provided by [1]. This data set contains 92 videos of different people performing following 10 actions: running, walking, skipping, jumping-jacks, jumping forward on two legs, jumping in place on two legs, jumping sideways, waving with two hands and waving with one hand. This data set is a

popular public benchmark used in many action recognition papers. The word vocabulary is computed using a subset of videos followed by generation of the signatures for each video. Testing is performed by the leave-one-out method. Since we have 10 actions, we trained 10 SVMs for each leave-one-out iteration using all the examples except the one on which testing was to be performed. The confusion matrix for this experiment is shown in Figure 5. It can be observed that we have obtained reasonable performance on most of the actions except 'jump' and 'skip'. These two actions are very similar to each other in the way that the actors bounce across the video. This will result in many similar 3D spatio-temporal cubes and therefore 3D SIFT descriptors.

In order to test the benefit of using our proposed 3D SIFT descriptor, we tested the performance against three other representations on the same data set. In the first case we used the 2D SIFT descriptor to describe single image based interest points. In the second case, independent 2D SIFT descriptors are computed around an interest point in consecutive frames to describe the spatio-temporal region around the point. The third representation is the spatio-temporal neighborhood of gradient magnitude used by [10]. The same experimental setup is used for all representations as was described previously. The performance is reported in Table 1. It can be seen that performance using our proposed 3D SIFT descriptors exceeds the other representations. The improvement over the two other descriptors which used information from a 3D neighborhood is especially striking. We believe the reason for this is that our descriptors are capturing the vital temporal information in a way which is missed by other representations.

Descriptor	Average Precision
2D SIFT	30.4%
Multiple 2D SIFT	47.8%
Gradient Magnitude	67.4%
3D SIFT	82.6%

Table 1: This table shows the average precision for different descriptors on the entire action data set.

5.1 Performance

SIFT has been popular in real time applications [6]. Currently in an un-optimized MATLAB implementation, a single full 3D SIFT descriptor is calculated in approximately 0.22 seconds. Conversion to C, or another compiled language, should yield a significant speedup. In our experiments 200 points were randomly selected from each video and used to create that video's 'signature'. Since SVM models can be trained offline, all that needs to be done for a given test video is random sampling and computation of the 3D SIFT features, matching the representative words for those features, and calculating the dot product for each possible action to find the final classification.

6. CONCLUSION

In this paper we have proposed an 3D SIFT descriptor and have demonstrated its improved performance on the task of action recognition in a bag of words paradigm. We have also performed comparative analysis of the performance of our proposed extension with other descriptors that are often used to describe spatio-temporal data. The results have

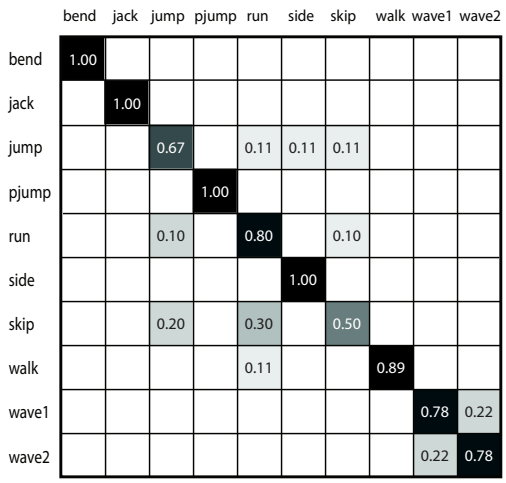

Figure 5: Confusion matrix demonstrating the performance of our method on the task of action classification.

demonstrated that our descriptor is able to outperform the existing descriptors on a publicly available action data set. In addition, we exploited co-occurrence based relationships between the words of a vocabulary to build more discriminative grouping histograms to represent a video. Future directions include applications to event detection and recognizing actions that are performed at different rates.

7. REFERENCES

[1] M. Blank et al., "Actions as Space-Time Shapes," *ICCV*, 2005.

[2] M. Brown er al., "Recognising Panoramas," *ICCV*, 2003.

[3] G. Csurka et al., "Visual Categorization with Bags of Keypoints", *ECCV*, 2004.

[4] A. Efros et al., "Recognizing Action at a Distance," *ICCV*, 2003.

[5] Y. Ke et al., "Efficient Visual Event Detection using Volumetric Features", *ICCV*, 2005.

[6] M. Lalonde et al., "Real-time eye blink detection with GPU-based SIFT tracking," *Fourth Canadian Conference on Computer and Robot Vision*, 2007.

[7] S. Lazebnik et al., "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *CVPR*, 2005.

[8] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, 2004.

[9] D. G. Lowe, "Object recognition from local scale-invariant features," *ICCV*, 1999.

[10] J. C. Niebles et al., "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *BMVC*, 2006.

[11] S. Se et al., "Vision-based Mobile Robot Localization And Mapping using Scale-Invariant Features," *International Journal of Robotics Research*, 2002.

[12] A. Yilmaz et al., "Actions Sketch: A Novel Action Representation," *CVPR*, 2005.