ACTION RECOGNITION USING SPATIO-TEMPORAL REGULARITY BASED FEATURES

Taylor Goodhart, Pingkun Yan, Mubarak Shah

School of Electrical Engineering & Computer Science,
University of Central Florida
http://www.cs.ucf.edu/~vision/

ABSTRACT

In this paper, a novel feature for capturing information in a spatio-temporal volume based on regularity flow is presented for action recognition. The regularity flow describes the direction of least intensity change within a spatio-temporal volume. Our feature consists of weighted histograms of the computed regularity flow around selected interest points. We then apply this new feature to recognizing actions with experiments on known benchmark dataset. A more discriminating representation of spatio-temporal volume is obtained by using the feature descriptors with the bag of words model. Action recognition is performed by using this new representation with a trained support vector machine. We show that by utilizing regularity flow based features, recognition can be performed with better performance than the best known features. Additionally, results suggest that our descriptor captures information otherwise not harnessed by existing methods.

Index Terms— Feature extraction, video analysis, regularity flow, action recognition.

1. INTRODUCTION

With the ever increasing amount of video information available, the problem of video content analysis is becoming increasingly important. It is a problem fraught with difficulties due to motion, however, including changes in perspective, lighting conditions, and scale. To complicate the issue, the variability between different actions is generally quite subtle.

Existing action recognition methods can be divided into two categories: model-based and feature-based. Model-based approaches generally resort either fitting a predefined structure - traditionally a human skeleton - to a video volume, or matching against predefined motion models [1]. These approaches perform well, but are constrained by the fact that explicit anthropometric models are required. Feature-based approaches are inherently more general - examining raw pixel data - at the expense of higher sensitivity to noise. Existing feature-based approaches have been designed to detect features such as optical flow, spatio-temporal corners [2], 3D SIFT [3], and high entropy regions [4]. A common trend among such detectors is that they are generalizations of existing 2D object recognition techniques. This is a natural extension, considering that both object recognition and action recognition face similar problems of occlusion, changes in perspective, multiple scales, and varying lighting conditions. Nevertheless, the fundamental difference between spatial domain and temporal domain is not fully considered.

Fig. 1. Some examples of typical actions: Bend, Jumping Jack, Jump, In Place Jump, Run, Side Step, Skip, Walk, One Handed Wave, Two Handed Wave.

Fig. 2. SPREF vectors computed from the spatio-temporal volume.

Our proposed method addresses the problem of action recognition using a feature-based approach; specifically, spatio-temporal regularity based feature (STRF). In our work, the spatio-temporal regularity is obtained by computing the spatio-temporal regularity flow (SPREF) [5], which is extracted by minimizing the energy of a spline curve approximation of the regularity flow over the entire volume. We also propose an interest point according to the distribution of the regularity flow. These feature descriptors are placed into a bag of words system in order to create a more robust description of video volumes, which is used in a support vector machine system in order to perform action recognition.

Our method is novel in that it considers a spatio-temporal video volume as a homogeneous unit to a far greater degree than existing methods, such as optical flow. This feature is suitable for action recognition as movement naturally creates highly regular patterns within video. Spatio-temporal regularity is also different from the popular SIFT [6] in that it uses video regularity, versus image gradient, as a source of infor-

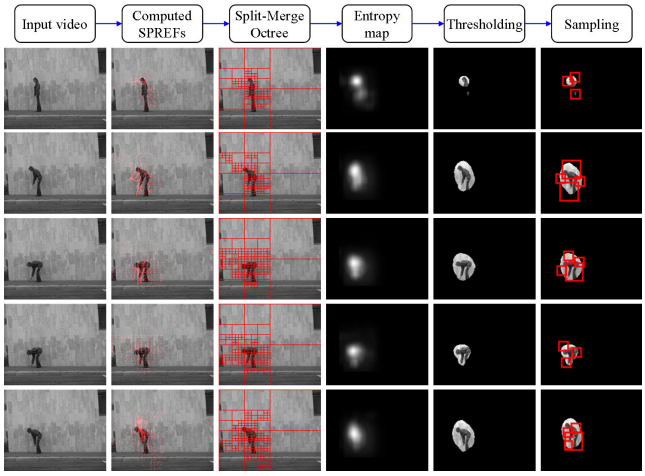

Fig. 3. Flow diagram outlining stages of feature extraction algorithm.

mation. We show that better results can be obtained on the well known Weizmann action dataset [7] using the proposed method compared to existing methods based on other features (see Fig. 1 for example frames from the dataset).

2. SPATIO-TEMPORAL REGULARITY BASED FEATURE (STRF)

2.1. Spatio-Temporal Regularity Flow (SPREF)

The spatio-temporal regularity flow - henceforth SPREF - is a $3D$ vector field describing the direction along which the intensity I of a video volume changes the least. We approximate the SPREF ($\mathcal{F}$) with a spline curve approximation that minimizes the energy E defined as,

$$E(\mathcal{F}) = \int_{\Omega} \left| \frac{\partial (I \star G)(x,y,t)}{\partial \mathcal{F}(x,y,t)} \right|^2 dx\, dy\, dt \qquad (1)$$

where G is a Gaussian filter. All motion is approximated as translational motion perpendicular to a propagation direction [5]. In our case, the propagation is the temporal axis.

The suitable scale of each sub-volume is determined by splitting the spatio-temporal volume using an oct-tree structure followed by a merge operation to remove redundant branches. For efficiency, however, we limit the depth of the octtree to 5 levels, for a maximum of 8^5 nodes. In the oct-tree, each node corresponds to a region of the volume. The child of a node corresponds to further subdivisions. We calculate the SPREF for each node and record the error, which is used as the criterion for further split or merge. An example of the SPREF is shown in Figure 2. For an in-depth discussion of the SPREF, we refer the readers to [5].

2.2. Feature Point Selection

In order to classify a spatio-temporal volume, we must efficiently encode the information provided by the regularity flow. Great success has been achieved by representing a spatio-temporal volume as a collection of interest points. We utilize this approach, basing our selection method on the entropy of angle histograms. Inspired by the work of Kadir and Brady [8], we locate interest points within a spatio-temporal volume by computing the entropy at an interest point and scale by examining a window about the point at each level of a scale pyramid. In practice, we quantize the SPREF vector angles into 8 bins and construct a histogram weighted by magnitude for a $16 \times 16 \times 16$ window. The entropy E of the interest point is calculated as

$$E(p,s) = -\sum_{a=1}^{8} h_s(p,a) \cdot log_2(h_s(p,a)) \qquad (2)$$

where $h_s(p,a)$ is the height of the a^{th} bin in the histogram for point p and s is the scale. We refer to the scalar field containing the entropy for each point in the spatio-temporal volume the entropy map. The entropy map has the physical interpretation of a region of highly regular movement; this is expected, as these are the regions with the highest degree of motion regularity. Regions of other types of movement - such as rotational movement - elicit a response, albeit not as strong. In contrast, corner-based approaches tend to generate interest points at joints as well as points of abrupt change in direction, such as the zenith of the path of a bouncing ball.

By thresholding this entropy map, we obtain a region of interest around points of high regularity. Within this region, interest points are selected by random. Traditionally, interest point selection methods search for extrema in a feature space. Our approach is novel in our use of interest points randomly selected within a thresholded region. This allows us to obtain a more representative set of interest points than would be possible by simply randomly selecting throughout the entire spatio-temporal volume. This approach also allows us to parameterize the number interest points selected, versus being limited by the number of extrema that may or may not exist.

2.3. Scale Selection

As the actors in a video may occur at different scales, representations of the SPREF need to be constructed at different scales. To achieve this, we separate the SPREF vector field of the original spatio-temporal volume into its component X and Y parts, resulting in two scalar fields. We then construct a scale pyramid by convolving these scalar fields with a Gaussian and down-sampling. The Gaussian acts as a low-pass filter in frequency space, where the low frequency components of the scalar fields correspond to information at larger scales. With an appropriate choice of Gaussian parameters, the new scalar fields will have half the bandwidth of the original representation. According to the Nyquist Shannon sampling theorem, we can sample the field at half the frequency and still retain all information. In practice this amounts to down-sampling the scalar fields. The scalar fields are then recombined to form a representation of the SPREF vector field

at a larger scale. There is the added computational benefit in that the new representation is now 1/4 the size of the original, which results in faster processing.

Note that no scaling is performed in the temporal dimension. Temporal scaling has a different physical interpretation than spacial scaling; specifically it corresponds to faster movement or a higher frame rate. In typical applications, the framerate and speed of actions in question are known *a priori*. As such, it is not necessary to perform scale selection in the temporal direction.

We compute an entropy map for each tier in the scale pyramid, and take the maximum at each point to create an entropy map representative of all scales. In practice this merging results in the entropy map appearing as a 'haze' or 'aura' about regions of high entropy. The gradual change in the intensity of the entropy map is a desirable property, as it affords us much control over the size of the region of interest generated by thresholding the entropy map.

2.4. Feature Descriptor

Having obtained interest points in a spatio-temporal volume, the next stage is the feature descriptor computation. Inspired by the success of histograms of gradients such as Lowe's SIFT descriptor [6], we propose a $16 \times 16 \times 16$ window about each interest point at the optimal scale. To place more emphasis on SPREF vectors closer to the interest point, we convolve the region with a Gaussian. The region thus has a spherical shape. In order to achieve rotation invariance, we compute the dominant SPREF vector angles for the region, and rotate the region accordingly. To protect against changes in light intensity, we normalize the magnitude of the SPREF vectors in the region. Finally we divide the region into $4 \times 4 \times 4$ subregions, and construct an angle histogram for each, weighted by the vector magnitudes. This results in a 512 dimensional feature vector for each interest point composed by 64 subregions with 8 bins. An example of this can be seen in figure 4.

This approach is similar to that of the SIFT and 3D SIFT, but different in three regards. First, vectors are derived from the regularity flow of a spatio-temporal volume, in contrast with the gradient. In a sense, our approach is mutually exclusive with the 3D SIFT. Second, the region about an interest point is rotated by the dominant SPREF angle for each frame, as opposed to the dominant gradient. Finally, only one rotation is performed in the XY plane, as opposed to the 3D SIFT, which performs two rotations. This is sufficient, because rotation in the temporal dimension has a different interpretation than that of the spatial dimensions.

2.5. Action Recognition

One observation is that our feature descriptor is not particularity well suited for a support vector machine (SVM) classifier due to its high dimensionality. Moreover, noise may create interest points that do not contain meaningful information, skewing results. Thus in order to recognize an action in a video, we do not use the feature vector descriptors directly but rather - as a preprocessing step - we utilize the *Bag of Words* approach to represent actions [9, 10]. First, the feature descriptors for an entire dataset are clustered according to a hierarchical K-means clustering algorithm. The resulting cluster centers then become our *Words*; the set of all words our *Vocabulary*. We then associate every feature descriptor in a video to the closest word. A video is now represented as a set of words. We then generate a word frequency histogram, called a *Signature*, for each video. While this frequency histogram could be used itself as a feature for our SVM system, there may be instances where particular words have a high co-occurrence rate, and thus contain similar information content. By merging these words into a new unit, we can reduce the dimensionality of our word frequency histogram - now a grouping frequency histogram - while maintaining all relevant information and as such obtain better results. In sum, by utilizing hierarchical K-means clustering in a bag of words framework, we are able to refine our original feature descriptors into a more representative descriptor.

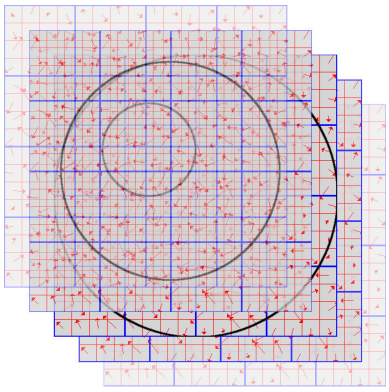

Fig. 4. An example of the region about an interest point. SPREF vectors are marked in red. Subregions are divided by the blue grid. The rings delineate the spherical shape of the region of interest. The descriptor we use spans across 16 slices at the optimal scale.

To test and train, we utilize the 'leave one out' strategy. This is done by selecting a video from the dataset and leaving it out of the training system. The remaining videos in the same action class are used as positive examples, and all other videos used as negative examples. To calculate an average, we iteratively leave out each video. In this manner, we are able to protect against biases that may arise from using any single video.

3. RESULTS

We conducted our tests using the Weizmann action dataset [7], which has become a standard test dataset for similar action recognition methods. We compare our results to Niebles's gradient magnitude based method [10] and Scov-

Table 1. Results of action recognition methods using different features for Irani dataset.

Gradient Magnitude[10]	72.8%
3D SIFT[3]	82.6%
STRF	**83.7%**

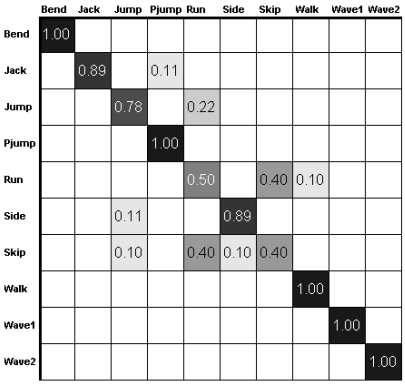

Fig. 5. Confusion Matrix. Left axis is actual action. Top axis is predicted action.

anner's 3D-SIFT based method [3]. We utilize the same classification framework as the 3D-SIFT based method, and thus our results wholly reflect the difference between the feature descriptors. The dataset consists of ten action classes such as jumping, walking, waving, and skipping, with each action being performed by nine actors, except for 'run' and 'skip', which have ten. In total, the dataset contains 92 videos. The results demonstrate that our performance is better than other known methods, as shown in table 1. Significantly, our approach perfectly recognize the actions of 'Walk1' and 'Walk2', in contrast with all other approaches. Fig. 5 shows the confusion matrix of the results.

After exploring the parameter space, we found that by utilizing 600 sample points were spatio-temporal volume, we obtained optimal results. For the sequences in which the actor crosses the entire field of view - 'jump', 'run', 'side', 'skip', and 'walk' - interest points at the next highest scale were more common, corresponding to a window size of $32 \times 32 \times 16$ pixels. For the hierarchical K-means grouping, the number of clusters k is set as 1000. This is a reasonable value, as it will generate words corresponding to an average of 55.2 feature descriptors. The initial positions of the cluster centers are chosen at random. This introduces a marginal degree of variability into the algorithm. Our best attained precision, shown in table 1, was 83.7% and the average precision was 79.4%.

Overall the results are quite promising. The majority of the action classes were correctly classified. In particular, the 'Bend', 'Pjump', 'Walk', 'Wave1', 'Wave2' actions were classified with 100% accuracy, as shown by figure 5. These actions were similar in that the actor was planted in one position, or moving relatively slowly. After examining the instances where the STRF did not perform as well, we discovered that this was due to a mixture of scaling and entropy-based thresholding. The misclassified actions were those in which the actor crosses the entire field of view. In particular, the majority of the voxels within the volume are all moving in a constant translational fashion in these cases. This type of motion generates a high response in our feature descriptor, due to the large degree of regularity. The largest response occurred at larger scales, which is to be expected, as the larger scales encompass more of the actor and less noise. The distinguishing aspects of each motion, however, occur at the smaller scales with lower entropy that were not selected. Colloquially, the translational movement of the actor 'drowned out' the differentiating aspects of the actions. This phenomenon did not occur in the correctly classified actions, as the scale of highest entropy and the scale of the differentiating aspects were the same.

4. CONCLUSION

In this paper we have proposed a novel new feature for action recognition: the spatio-temporal regularity flow (STRF). We have compared our results to those of others on a known database, and shown results that perform as well as the best feature-based systems. More importantly, our feature was able to classify certain actions better than other methods, suggesting that our feature encodes unique information. We hope to show in future works that the STRF can be used in conjunction with alternative formulations of the STRF to obtain improved results.

5. REFERENCES

[1] D.M. Gavrila and L.S. Davis, "3D model-based tracking of humans in action: a multi-view approach," in *CVPR*, 1996.

[2] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*, 2003.

[3] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT descriptor and its application to action recognition," in *ACM Conf. Multimedia*, 2007.

[4] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *IEEE Trans. SMC*, vol. 36, pp. 710–719, 2006.

[5] O. Alatas, P. Yan, and M. Shah, "Spatiotemoral regularity flow (SPREF): Its estimation and applications," *IEEE Trans. CSVT*, vol. 17, no. 5, pp. 584–589, May 2007.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[7] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *ICCV*, 2005.

[8] T. Kadir and M. Brady, "Scale saliency : A novel approach to salient feature and scale selection," in *CVIE*, 2003.

[9] G. Csurka et al., "Visual categorization with bags of keypoints," in *ECCV*, 2004.

[10] L. Fei-Fei et al., "Unsupervised learning of human action categories using spatial-temporal words," in *BMVC*, 2006.