Automated detection of fighting styles using localized action features

Ankush Aniket Mishra
PES Center for Pattern Recognition
Dept. of Information Science and Engineering,
PES Institute of Technology Bangalore South Campus
Bangalore, India - 560100
Email: ankushmishra9@gmail.com

Gowri Srinivasa
PES Center for Pattern Recognition
Dept. of Computer Science and Engineering,
PES Institute of Technology Bangalore South Campus
Bangalore, India - 560100
Email: gsrinivasa@pes.edu

***Abstract*—In this paper, we propose a recognition method for the classification of martial arts videos. In our approach, we utilize the spatio-temporal interest points, to detect regions associated with movement in a sequence of frames in the videos. This is then used to construct a bag of words for all the descriptors produced. As a part of the bag of words, we cluster the all video descriptors and then histogram each video as a representation of these clusters. This representation is what we input to two classifiers, kNearestNeighour and Support Vector Machine and the results obtained from them is promising to be used for further applications.**

***Index Terms*—Computer Vision, Actions, Spatio-Temporal Interest Points**

I. Introduction

The problem of making the computer see and understand the human environment remains at the crux of computer vision. Given a video to the computer, can we ask it to understand what is happening in the video? Can we find other videos represented by similar movements? Can this process be automated? These types of fundamental questions is what we attempt to address in this work.

Action recognition remains as one of the more fundamental problems of computer vision. The ability to determine if a given video contains some action, is useful in multiple use case scenarios. In our work, we present a methodology for recognition of martial arts videos based on leveraging the power of spatio-temporal interest points and the use of supervised learning. Once our model is constructed, we can use this as a precursor to a tagging system for the purpose of efficient information retrieval. This is extremely helpful for first time learners where in a larger design we can classify videos based on their fighting styles and discrete stances.

The recognition of martial arts videos themselves represent a large subset of problems which occurs with other approaches, such as the existence and visibility of multiple persons present in the sequence, along with that there is also the consideration of the camera movement which causes some immobile aspects to come up as mobile parts. There's also the consideration that, with respect to actions we can have a discrete set of sequences to be considered as actions, for example, walking. But in our case, it being martial arts, a lot of times we find some unorthodox moves occurring and a lot of moves that occur are into a direct flowing transition to another movement. This causes a lot of doubt on what can be considered a sequence or not.

In our problem, we take the traditional approach similar to that of Schuldt et al. [1] and, Niebles and Fei-Fei [2], where we perform our classification task in three steps: (1) **pre-processing**, where we first crop and segment our respective dataset, to clean out the unnecessary aspects of the videos (2) **feature extraction**, we utilize the spatio-temporal interest points as defined by Laptev [3], as a means of detection of points, then we use histogram of gradients and histogram of optical flow as our descriptors for our points. (3) **classification**, in this stage, we construct a training vocabulary made of all the descriptors, then cluster them. After clustering, we then construct the representation of each video as a histogram of these clusters. Beyond which, we apply either Support Vector Machines or kNearestneighour as classifiers to our videos.

The use of local features of optical flow and space-time gradients over a global representation of these features, is due to the knowledge that the global representations will also be influenced by dynamic backgrounds, motions of multiple events simultaneously, over discrete events in the video. The local features allow us the ability to utilize the aforementioned discrete events in a sequence which makes it easier to remove any superfluous information.

II. Background and Related Work

Human action recognition comes with a large pedigree of work being accomplished over time by numerous researchers in the field. [4].

An early approach by Bobick and Davis [5], was the use of global representations of the video by utilizing *motion history images* and *motion energy images* to create temporal templates, this was then used along with a matching algorithm. A similar approach by Blank et al. [6], along with producing the Weizmann human action dataset, utilize the properties of space-time shapes to produce space-time features like space-time saliency, orientations, and with global features. This method, relies on the background remaining static, due to the dependency on the separation between foreground and background.

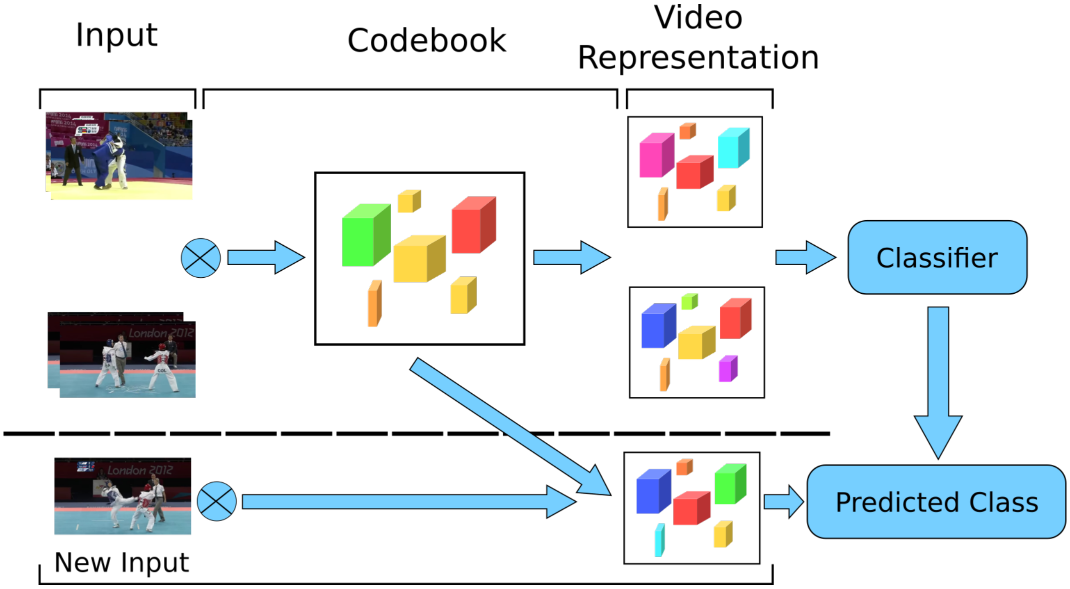


Fig. 1. This image represents the complete working of the procedure that we follow to find the solution to our problem.

Efros et al. [7] utilize the localized optical flow measurements to compare corresponding spatio-temporal volumes for person-centered in low resolution sequences to represent motion. These volumes are then compared against a database of annotated video sequences.

Schetman and Irani [8], introduce a behaviour based similarity measure based on the underlying motions field. They do this, utilizing a action template, which will be correlated in the videos, and at each pixel a need to calculate gradients of the video patch, thus requiring a significant computation to deal with the task. Their work, albeit has the advantage of not requiring to perform foreground/background segmentation, but the computation requirement puts that down.

Other approaches like Boiman and Irani [9], take an approach of representing a sequence as an ensemble of local spatial patches to match a video query against a set of query videos in the database. Dense sampling is a necessity in their approach, which makes this computationally expensive. Niebles and Fei-Fei [2], propose an unsupervised learning method, by utilizing probabilistic latent semantic analysis (pLSA), a generative model. They utilize a bag of words assumption, and produce models which learn the action category based on their histograms.

Laptev [3] presents an interest point detector based on Harris's interest point detectors, which detect localized features which have large variations, with respect to both space and time. Dollár et al., [10] propose a spatio-temporal interest point, based as a set of separable linearly filters, producing a larger number of detections.

Once, these interest points have been extracted they have been utilized in various ways. Schuldt et al. [1], Dollár et al. [10], Oikonomopoulos et al. [11] and the rest have used them with classifiers to learn and understand actions.

Along with these traditional approaches, some recent advancements were made by the use of Convolutional Neural Networks(CNN) to the action recognition and classification problem. Kaparthy et al. [12] utilize CNN's on large scale video classification on a dataset of 1 Million Youtube videos, belonging to 487 classes, while taking advantage of the local spatio-temporal information. Shikhar Sharma et al. [13] also shows the use of Recurrent neural networks with Long Short-Term Memory units, to selectively learn relevant features of the videos.

As can be noted in some approaches, assumptions and dependencies on static background, stationary camera, lack of any zooming or camera movement, and limited or simplistic movement provides a very restricted solution to this vast problem. Also, videos containing multiple persons present in the same frame along with extended dynamic movement is something that is scarcely found in previously researched approaches.

We present our approach to show the handling of data consisting of moving cameras, dynamic movements and fast reflexes, as a consequence of the actions themselves. Some approaches depend on the separation of the foreground and background, contrast this to our approach where we attempt to solve this by using the HOG and HOF features of the localized patches. The choice of the local features over global features is another significant point on which we decide improve on due to their independent nature towards dynamic backgrounds, while also being the compact and abstract representation of the actions presented.

III. Representation

The complete representative workflow of the process is as shown in Figure 1. To represent each video's motions we use the local space time features, as defined by Laptev [3] and the same is used by Schuldt et al. [1]. The local features we detect

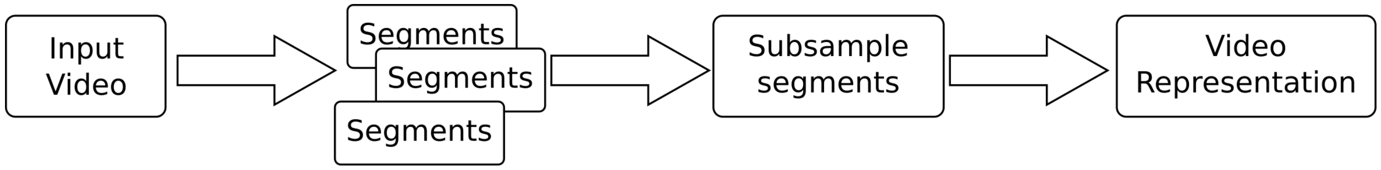


Fig. 2. Preprocessing workflow

is essentially a modified Harris detector for the 3 dimensional plane, with time being the third dimension. For detecting local features in an image sequence $f(x, y, t)$ we construct it's scale space representation

$$L(\cdot, \sigma^2, \tau^2) = f * g(\cdot, \sigma^2, \tau^2) \tag{1}$$

where g is the Gaussian convolution kernel, given by:

$$exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)/\sqrt{(2\pi)^3\sigma_l^4\tau_l^2} \tag{2}$$

This is then used to compute the second-moment matrix using image gradients:

$$\nabla L = (L_x, L_y, L_t)^T \tag{3}$$

within the Gaussian neighbourhood of each point

$$\mu(\cdot, \sigma^2, \tau^2) = g(\cdot, s\sigma^2, s\tau^2) * (\nabla L(\nabla L)^T) \tag{4}$$

and define position of features by local maxima of $H = det(\mu) - ktrace^3(\mu)$ over (x, y, t). The spatio-temporal neighbourhood of the neighbourhood of features in space and time is then defined by spatial and temporal parameters (σ, τ) of the associated Gaussian kernel. As shown by Laptev, the size of features can be adapted to the velocity of local pattern by automatically selecting scales parameters (σ, τ). Also, the shape of the features can be adapted to the velocity of the pattern, making the features work over camera movement.

Spatio-temporal neighbourhoods of local features contain information about the motion and spatial appearance of events in the sequences. To capture this information, we compute the spatio-temporal jets as

$$l = (L_x, L_y, L_t, L_{xx}, \ldots, L_{tttt}) \tag{5}$$

at the center of each feature using normalized derivatives $L_{x^m y^n t^k} = \sigma^{m+n}\tau^k(\partial_{x^m y^n t^k} g) * f$ computed using selected scales values for the parameters (σ^2, τ^2). To enable invariance with respect to relative camera movement, we also warp the neighbourhoods of features using estimated velocity prior to computation of l.

K-means clustering of descriptors l in the training set gives a vocabulary of events h_i. The numbers of features with labels h_i, in a particular sequence defines a feature histogram $H = (h_1, \ldots, h_n)$. We use these histograms as the representation for the videos, when recognizing actions in them.

The extracted Spatio-temporal points, can be seen in Figure 3

IV. Video Classification

For classification, we utilize both SVM's and k-Nearest neighbour classifiers.

Support Vector Machines are large margin classifiers where they follow the Structural Risk Minimization principle and along with that, it also provides high generalization ability [14]. They are universal learners, and by the use of a 'kernel trick' they can be used to learn polynomial classifiers, RBF's and neural networks. Another feature of SVM's is that they allow capacity control, which makes it possible to take into account the amount of training data.

Consider the training data as a set of m samples x_i and labels y_i associated with them, given by $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_m, y_m)$

where $x_i \in \Re^N$ is a feature vector and $y_i \in 0, 1$ it's class label.

Given a hyperplane $w \cdot x + b = 0$ in some space H, then the goal in training a Support Vector Machine is to find the separating hyperplane with the optimal margin.

For training the SVM, the first step is to choose non-linear ψ-functions that map the input to a higher dimensional space [15], with the help of a Kernel, $K(x, x')$, given by

$$K(x, x') = \sum_i \psi(x)\psi(x')$$

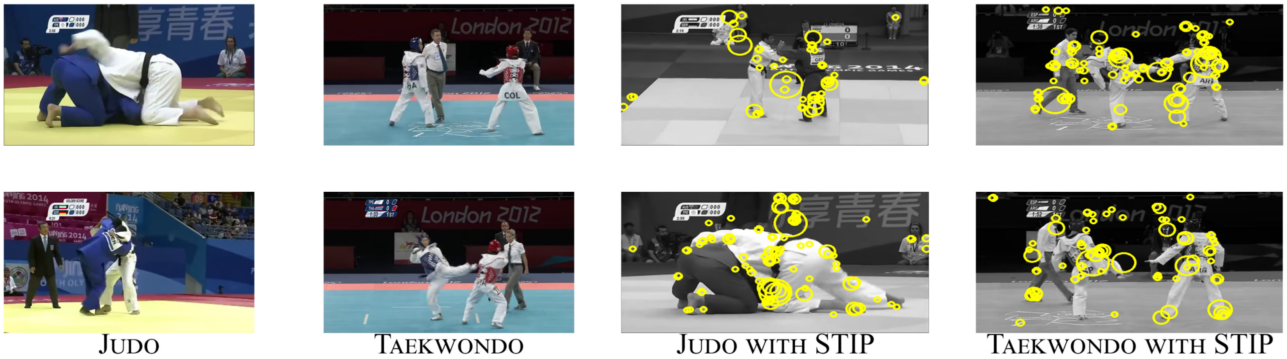
The support vectors are the training samples that define the optimal separating hyperplane and are the most difficult patterns to classify. The optimal values of w and b can be found by solving a constrained minimization by utilizing Lagrangian Multipliers to solve by duality produced when calculating the support vectors.

$$f(x) = sgn(\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b)$$

where α_i and b are found by the learning algorithm.

V. Experimental Results

For our dataset we use the Olympics as a source of videos, including the 2012 London Olympics and the 2014 Nanjing Youth Olympics. We select two major classes of videos, Judo and Taekwondo, as them being Olympic sports affords them more video output for the general audience.



Judo

Taekwondo

Judo with STIP

Taekwondo with STIP

Fig. 3. Samples from the dataset, along with Spatio-temporal Interest points detected from the videos

A. *Preprocessing*

The overall preprocessing workflow can be seen in Figure 2. For the preprocessing stage, we partitioned the videos into sequences based on two aspects:

- Change in camera angle, this allowed the video to have ubiquitous camera position without any additional changes in the video itself.
- Referee signals, this allowed no extraneous movement to be captured, as most of the unnecessary movements occur during after the signal.

We also, made sure that no part of the video included walking at any point of time in the video. Some videos featured audiences in the background, most of these videos were removed as they caused too much noise during the calculation of the interest points, whereby a lot of superfluous points were present in the descriptors.

Videos based on different viewpoints, specifically side and top views, were also used to provide different viewing angles. Though, a large majority were only available as side views.

Some sequences which were longer than the average length were also dissected into smaller segments to make them more feasible due to the shorter lengths.

After segmenting, we also cropped some of these videos to reduce extra noise, from the videos.

These, videos after segmenting and cropping, were then sampled to include only every 3rd frame in the videos. This allowed for faster overall computation without negligible loss in accuracy.

In the end we had about 150 sequences of Judo, and Taekwondo respectively. Almost all sequences had a referee present either in the background or in the foreground near the contestants. Along with that, each video was a fighting sequence amongst two participants, over a single person as seen in general action recognition datasets like KTH [1]. Also, the backgrounds were not static due to the fact that we had a moving camera. Each video was 640×360 pixels and with a 25fps.

We had about 8 Olympic matches worth of videos per class, with unique persons per match. The resultant video samples, are as shown in Figure 3.

B. *Feature Extraction*

As stated earlier we utilized spatio-temporal interest points to calculate the features. We used the Harris3D detector, along with hoghof as the descriptors, similar to what Schuldt et al. [1] utilized. We also utilized σ^2 was selected as 4 and τ^2 as 2, the spatial patch size for the descriptors was set to 9 and temporal patch size 4. The descriptors are of size 160 representing both hog and hof features per point.

Once the descriptors have been calculated we utilize a cluster size of 1500 to create a codebook based on the Bag of Words, and then each image as stated earlier is represented as a histogram of these clusters.

C. *Classification*

For classification, we utilize two different classifiers,

- SVM with either a linear kernel and a RBF kernel.
- Nearest Neighbour classifier

For cross validation we ran a 10 K-Fold Cross Validation for the testing and optimizing the hyperparameters.

D. *Results*

As shown in the Table I, which displays the average accuracy of the classifiers, along with the proportion of the true positives present per class. It is visible that the Linear SVM gives the best performance for all training set in our k-fold cross validation, with an average accuracy of 91.6646% and standard deviation of 6.4336%, whereas the Nearest Neighbour performs at an accuracy of 86.4876% with a standard deviation of 7.3799%. The SVM with the RBF kernel performs the worst, at 59.0109% and 9.2662%. The SVM outperforms the Nearest Neighbour classifier as expected.

TABLE I
ACCURACY OF DIFFERENT CLASSIFIERS AND PER CLASS TRUE POSITIVES.

	kNN	SVM	SVM + RBF
Average Accuracy	$86.48\% \pm 7.38\%$	$91.66\% \pm 6.43\%$	$59.01\% \pm 9.26\%$
Judo (True Positives)	$79.06\% \pm 11.88\%$	$89.23\% \pm 8.58\%$	$94.95\% \pm 5.6\%$
Taekwondo (True Positives)	$94.12\% \pm 7.37\%$	$94.23\% \pm 7.01\pm$	$22.47\% \pm 17.03\%$

VI. DISCUSSION

There are a lot of things to consider in our work that makes this problem especially challenging to work on while also achieving high accuracy. Firstly, in these videos, there are always multiple people present in the same frame causing confusion as to which part of the video is contains the actual movement. For example, in some videos, the referee's movement is not separable from the video during preprocessing. The movement of the cameras is presents another major issue where it renders a part of the immobile background as mobile, thus in some cases causing erroneous points to be detected.

Another issue to discuss is that, the length of a sequence defined in a video, is currently based on the change of camera angle or the referee's stopping signal, but the sequence could be practically anything, where we could consider even each discrete movement as a sequence. This determination of sequence length, is a major part of what we intend to achieve.

Unlike the other datasets, like the KTH [1] or Weizmann, where we have a discrete set of actions within a small time frame. In ours, because it is a martial arts dataset, it renders a lot more interest points to be detected due to the large amount of movement that occurs. This is different from say KTH, where any action has only a small amount of points being rendered.

On the selection of videos, we could have utilized training videos for each fighting style but alas, this was not feasible due to the smaller length of the videos, or that since, it's a training video, a large amount of time is spent just explaining the move.

Although we utilized STIP's based on Laptev [3] as a basis for our work, we could have used Dollár, et al. [10] work as well for the detector points based on cuboids, this could quite possibly have produced more dense and accurate representations.

This work could be extended to produce an automated martial arts tagging system for efficient information retrieval by an indexing system. We could use this to tag stances and movements of various martial arts, which would be helpful to new and old learners of the art.

VII. CONCLUSION

In this paper, we present our work on detection and classifying of martial arts videos, based on the use of bag of words representation of the videos with supervised learning. We utilized localized spatio-temporal interest points as our detectors, along with optical flow, and oriented gradients as the descriptors. The results we have obtained is promising, which leads us to think that it could form a precursor to a tagging and information retrieval system, for further efficient learning and use of martial arts videos.

REFERENCES

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36.

[2] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[3] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[4] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[5] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[6] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1395–1402.

[7] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings of the International Conference on Computer Vision (ICCV03)*. IEEE, 2003, p. 726.

[8] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 405–412.

[9] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International journal of computer vision*, vol. 74, no. 1, pp. 17–31, 2007.

[10] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[11] A. Oikonomopoulos, M. Pantic, and I. Patras, "Sparse b-spline polynomial descriptors for human activity recognition," *Image and vision computing*, vol. 27, no. 12, pp. 1814–1825, 2009.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[13] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[14] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[15] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.