Modified Time Flexible Kernel for Video Activity Recognition using Support Vector Machines

Ankit Sharma[1], Apurv Kumar[1], Sony Allappa[1], Veena Thenkanidiyoor[1], Dileep Aroor Dinesh[2] and Shikha Gupta[2]

[1]*National Institute of Technology Goa, India*

[2]*Indian Institute of Technology Mandi, Himachal Pradesh, India*

Keywords: Video Activity Recognition, Gaussian Mixture Model based Encoding, Support Vector Machine, Histogram Intersection Kernel, Hellinger Kernel, Time Flexible Kernel, Modified Time Flexible Kernel.

Abstract: Video activity recognition involves automatically assigning a activity label to a video. This is a challenging task due to the complex nature of video data. There exists many sub activities whose temporal order is important. For building an SVM-based activity recognizer it is necessary to use a suitable kernel that considers varying length temporal data corresponding to videos. In (Mario Rodriguez and Makris, 2016), a time flexible kernel (TFK) is proposed for matching a pair of videos by encoding a video into a sequence of bag of visual words (BOVW) vectors. The TFK involves matching every pair of BOVW vectors from a pair of videos using linear kernel. In this paper we propose modified TFK (MTFK) where better approaches to match a pair of BOVW vectors are explored. We propose to explore the use of frequency based kernels for matching a pair of BOVW vectors. We also propose an approach for encoding the videos using Gaussian mixture models based soft clustering technique. The effectiveness of the proposed approaches are studied using benchmark datasets.

1 INTRODUCTION

Activity recognition in video involves assigning an activity label to it. The task is easily carried out by human beings with little effort. However, it is a challenging task for a computer to automatically recognize activity in video. This is due to the complex nature of video data. There exists spatio temporal information in videos which is important for activity recognition. A video activity includes several sub-actions whose temporal ordering is very important. For example, 'getting-into-a-car' activity involves sub-activities like, 'standing near car door', 'opening the door' and 'sitting on car seat'. These three sub-actions should happen in the same order to recognize the activity as 'getting-into-a-car'. If the order of sub-actions is reversed, the video activity corresponds to 'getting-out-of-a-car'.

Video activity recognition involves extracting low-level descriptors that capture the spatio-temporal information existing in videos (Laptev, 2005; Paul Scovanner and Shah, 2007; Alexander Klaser and Schmid, 2008). Capturing spatio-temporal information in videos leads to represent a video as a sequence of feature vectors as $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$ where $\mathbf{x}_t \in \mathcal{R}^d$ and T is the length of the sequence. Since the videos are of different lengths, the lengths of the corresponding sequences of feature vectors also vary from one video to another. Hence, video activity recognition involves classification of varying length sequences of feature vectors. The approaches to video activity recognition require to consider temporal information in videos. Conventionally hidden Markov models (HMMs) (Junji Yamato and Ishii, 1992) are used for modelling sequential data such as video. The HMMs are statistical models that are built using non-discriminative learning based approaches. The activity recognition in video is a challenging task that needs special focus on discrimination among the various activity classes. Considering discrimination among the activity classes requires using the discriminative learning based approaches. Recently, support vector machines (SVMs) based classifiers are found to be very effective discriminative learning based classifiers for activity recognition (Dan Oneata and Schmid, 2013; Wang and Schmid, 2013; Adrien Gaidon and Schmid, 2012). In this work,

Sharma, A., Kumar, A., Allappa, S., Thenkanidiyoor, V., Dinesh, D. and Gupta, S.
Modified Time Flexible Kernel for Video Activity Recognition using Support Vector Machines.
DOI: 10.5220/0006595501330140
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 133-140
ISBN: 978-989-758-276-9

we explore an SVM-based approach for activity recognition in videos.

An SVM-based approach to video activity recognition should address the issue of handling varying length sequences of feature vectors. There are 2 approaches to address this issue. First one is by encoding the varying length data into a fixed length representation and then using standard kernels such as linear kernel or Gaussian kernel to build SVM-based activity recognizer. In this approach for video encoding the temporal information is lost. The second approach is to design kernels that can consider the varying length data. The kernels that consider varying length data are known as dynamic kernels (Dileep and Chandra Sekhar, 2013). A dynamic kernel named time flexible kernel (TFK) is proposed in (Mario Rodriguez and Makris, 2016) for SVM-based activity recognition in videos. In this approach, videos are encoded as sequences of bag-of-visual-words (BOVW) vectors. The visual words represent local semantic concepts and are obtained by clustering the low-level descriptors of all the videos of all the activity classes. The BOVW representation corresponds to the histogram of visual words that occur in an example. The low level descriptors extracted from a fixed size window of frames of a video is encoded into a BOVW vector. The window is moved by a set of frames and the next BOVW vector is obtained. In this way, videos are encoded into sequences of BOVW vectors. Matching a pair of videos using TFK involves matching every BOVW vector from a sequence with every BOVW vector from the other sequence. It is observed in (Mario Rodriguez and Makris, 2016) that in an activity video, middle of the video has the core of the activity. To consider this, the TFK uses weights for match score between a pair of BOVW vectors. The weight for matching the BOVW vectors from the middle of two sequences is higher than for matching a BOVW vector from the middle of one sequence and a BOVW vector from the end of the other sequence. The TFK uses linear kernel for matching a pair of BOVW vectors.

In this work, we propose to modify the TFK by exploring better approaches to match a pair of BOVW vectors. The BOVW representation corresponds to frequency of occurrence of visual words. It is shown in (Neeraj Sharma and A.D, 2016) that the frequency based kernels are suitable for matching a pair of frequency based vectors. Hence in this work, we propose modified TFK (MTFK) that uses frequency based kernels for matching a pair of BOVW vectors. An approach considered for encoding of a video into a sequence of BOVW vectors also plays an important role. In this work we propose to encode the video using a Gaussian mixture model (GMM) based approach. The GMM is a soft clustering technique and BOVW vectors are obtained using soft assignment of descriptors to clusters. Effectiveness of the proposed kernel is verified using the video activity detection on a benchmark dataset.

The paper makes the following contributions towards exploring activity recognition in videos using SVMs. First, we explore frequency based kernels to match a pair of BOVW vectors one each from two sequences and obtain modified TFK. This is in contrast to (Mario Rodriguez and Makris, 2016), where linear kernel is used to match a pair of BOVW vectors. Our second contribution is in exploring GMM-based soft clustering to encode a video into a sequence of BOVW vectors. This is in contrast to (Mario Rodriguez and Makris, 2016), where codebook based soft assignment is used for video encoding.

This paper is organized as follows: A brief overview of approaches to activity recognition is presented in Section 2. In Section 3, we present modified time flexible kernel. The experimental studies are presented in Section 4. In Section 5, we present the conclusions.

2 APPROACHES TO ACTIVITY RECOGNITION IN VIDEOS

For effective video activity recognition, it is necessary to use spatio-temporal information available in videos. This can be done by extracting features that capture spatio-temporal information and then using classification models that are capable of handling such data. Feature extraction plays an important role in any action recognition task. The feature descriptors must be capable of representing spatio-temporal information. It is also important for the feature descriptors to be invariant to noise (Shabou and LeBorgne, 2012; Dalal and Triggs, 2005; Jan C. van Gemert and Zisserman, 2010; Dalal et al., 2006). To consider temporal information, trajectory based feature descriptors are useful (Laptev, 2005; Paul Scovanner and Shah, 2007; Alexander Klaser and Schmid, 2008; Wang and Schmid, 2013; Adrien Gaidon and Schmid, 2012; Zhou and Fan, 2016; Orit Kliper-Gross and Wolf, 2012; Jagadeesh and Patil, 2016; Chatfield Ken and Andrew, 2011). This category of feature descriptors mainly focus on tracking specific regions or feature points over multiple video frames. Once the spatio-temporal feature descriptors are extracted, suitable classification models may be used for activity recognition in videos. Conventionally hidden Markov models (HMMs) (Junji Yamato and Ishii,

1992) are used for modelling sequential data such as video. The HMMs are statistical models that are constructed using non-discriminative learning based approaches. The activity recognition in video is a challenging task that needs special focus on discrimination among the various activity classes. Such discrimination among the activity classes is possible using the discriminative learning based approaches. In this direction, condition random fields (Jin Wang and Liu, 2011) are found useful in modelling the temporal information. Another popular category of classifiers that focus on discriminative learning are SVM-based classifiers. In (Jagadeesh and Patil, 2016; Zhou and Fan, 2016; Dan Xu and Wang, 2016; Cuiwei Liu and Jia, 2016), SVM-based human activity recognition is proposed. An important issue in SVM-based approach to activity recognition in videos is the selection of suitable kernels. Standard kernels such as Gaussian kernel, polynomial kernel etc., cannot be used for varying length video sequences. The kernels that consider varying length data are known as dynamic kernels (Dileep and Chandra Sekhar, 2013). A dynamic kernel named time flexible kernel (TFK) is proposed in (Mario Rodriguez and Makris, 2016) for SVM-based activity recognition in videos. The TFK involves weighted matching of every feature vector from a sequence of feature vector with every feature vector from the other sequence of feature vectors. In this paper we propose to modify the TFK by exploring better approach for matching a pair of feature vectors. We present the modified TFK in the next section.

3 ACTIVITY RECOGNITION IN VIDEOS USING MODIFIED TIME FLEXIBLE KERNEL BASED SUPPORT VECTOR MACHINES

The activity recognition in videos involves considering the sequence of feature vectors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$ representation of videos. Here, $\mathbf{x}_t \in \mathcal{R}^d$ and T corresponds to the length of the sequence. Every video undergoes an encoding process. Then a suitable kernel is computed to match a pair of activity videos to recognize the activity using support vector machines (SVMs). In this section, we first present approaches for video encoding. Then we present the proposed modified time flexible kernel for matching two activity videos.

3.1 Video Encoding

In this section, we present the process of encoding a video into a sequence of bag-of-visual-words (BOVW) vectors. The video activity is a very complex phenomenon that involves various sub-actions. For example activities such as 'getting-out-of-a-car' and 'getting-into-a-car' have many common sub-actions. For effective activity detection, the temporal ordering of the sub-activities is important. Hence it is important to retain the temporal information while encoding the video. To retain the temporal information, a video sequence $\mathbf{X}$ is split into parts by considering sliding window of a fixed number of frames. A video sequence $\mathbf{X}$ is now denoted as a sequence of split videos as $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^i, \ldots, \mathbf{X}^N)$, where $\mathbf{X}^i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it}, \ldots, \mathbf{x}_{iT_i})$. Here, N corresponds to the number of split videos, $\mathbf{X}^i$ is the sequence of feature vectors for ith split video whose length is T_i and $\mathbf{x}_{it} \in \mathcal{R}^d$. Every split video is encoded into a BOVW representation.The BOVW representation of a video segment $\mathbf{X}^i$ corresponds to a vector of frequencies of occurrence of visual words denoted by $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ik}, \ldots, y_{iK}]^\top$. Here, y_{ik} corresponds to the frequency of occurrence of the kth visual word in the video segment and K is the number of visual words. A visual word represents a specific semantic pattern shared by a group of low-level descriptors. The visual words are obtained by clustering all the d-dimensional low-level descriptors, $\mathbf{x}$ of all the videos. To encode $\mathbf{X}^i$ into $\mathbf{y}_i$, every low-level descriptor $\mathbf{x}$ of a video segment $\mathbf{X}^i$ is assigned to the closest visual word using a suitable measure of distance. In this work, we explore a soft clustering method such as Gaussian mixture model (GMM) to encode the videos. In the GMM-based method, the belongingness of a feature vector $\mathbf{x}$ to a cluster k is given by the posterior probability of that cluster. Then y_k is computed as $y_k = \sum_{t=1}^{T} \gamma_k(\mathbf{x}_t)$ where $\gamma_k(\mathbf{x}_t)$ denotes the posterior probability and T is the number of descriptors. This is in contrast to the codebook based soft assignment approach followed in (Mario Rodriguez and Makris, 2016). Encoding split videos into BOVW representation results in a sequence of BOVW representation for a video, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n, \ldots, \mathbf{y}_N)$ where $\mathbf{y}_n \in \mathcal{R}^K$.

In this work, we consider the following two video encoding approaches: 1)Encoding a video $\mathbf{X}$ as a sequence of BOVW representation $\mathbf{Y}$ and 2) Encoding entire video $\mathbf{X}$ as a BOVW representation $\mathbf{z}$. An important limitation of encoding entire video $\mathbf{X}$ into $\mathbf{z}$ is that the temporal information in the video is lost. The number of split videos differ from one video to

another as each videos are of different durations. This results in representing videos as varying length patterns when they are encoded using the first approach. For SVM-based activity recognition, it is necessary to consider a suitable kernel that matches the activities in the videos that are in the form of varying length patterns such as $\mathbf{Y}$. Commonly used standard kernels such as Gaussian kernel or polynomial kernel cannot be used for the sequences of feature vectors. Kernels that consider the varying length patterns are known as dynamic kernels (Dileep and Chandra Sekhar, 2013). In this work we propose a dynamic kernel, namely modified time flexible kernel for the SVM-based activity recognition in videos.

3.2 Modified Time Flexible Kernel

Let $\mathbf{Y}_i = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{in}, \ldots, \mathbf{y}_{iN}\}$ and $\mathbf{Y}_j = \{\mathbf{y}_{j1}, \mathbf{y}_{j2}, \ldots, \mathbf{y}_{jm}, \ldots, \mathbf{y}_{jM})$ correspond to the sequence of BOVW vectors representation of the videos $\mathbf{X}_i$ and $\mathbf{X}_j$ respectively. Here, N and M correspond to the number of split videos and $\mathbf{y}_{in}, \mathbf{y}_{jm} \in \mathcal{R}^K$. The time flexible kernel (TFK) involves matching every BOVW vector from $\mathbf{Y}_i$ with every BOVW vector from $\mathbf{Y}_j$. In (Mario Rodriguez and Makris, 2016), linear kernel (LK) is used for matching a pair of BOVW vectors. In an activity video, the core of activity happens in the center of the video. Hence to effectively match a pair of videos, it is necessary to ensure that their centers are aligned. This is achieved by using a weight, w_{nm} for matching between $\mathbf{y}_{in}$ and $\mathbf{y}_{jm}$. The value of w_{nm} is large when $n = N/2$ and $m = M/2$ ie., matching at the center of two sequences. The value of w_{nm} for $n = N/2, m = 1$ will be smaller than w_{nm} for $n = N/2, m = M/2$. The details of choosing w_{nm} can be found in (Mario Rodriguez and Makris, 2016). Effectively the TFK is a weighted summation kernel as given below:

$$K_{\text{TFK}}(\mathbf{Y}_i, \mathbf{Y}_j) = \sum_{n=1}^{N} \sum_{m=1}^{M} w_{nm} K_{\text{LK}}(\mathbf{y}_{in}, \mathbf{y}_{jm}) \quad (1)$$

Here, $K_{\text{LK}}(\mathbf{y}_{in}, \mathbf{y}_{jm})$ corresponds to linear kernel computed between $\mathbf{y}_{in}$ and $\mathbf{y}_{jm}$. In principle, any kernels on fixed-length representation can be used in place of $K_{\text{LK}}(.)$ in (1). The widely used non linear kernels on fixed-length representations such as the Gaussian kernel (GK) or the polynomial kernel (PK) can be used. An important issue in using GK or PK is that a lot of effort is needed to tune the kernel parameters. In this work, each split video is represented using BOVW representation which is a histogram vector representation. The frequency-based kernels for SVMs are found to be more effective when the examples are represented as non-negative vectors i.e. histogram vectors (Neeraj Sharma and A.D, 2016). The frequency-based kernels are also successfully used for image classification when each image is represented in BOVW representation (Chatfield Ken and Andrew, 2011). In this work, we propose to use two frequency-based kernels namely, histogram intersection kernel (HIK) (Jan C. van Gemert and Zisserman, 2010) and Hellinger's kernel (HK) (Chatfield Ken and Andrew, 2011) as non linear kernels to modify the TFK.

3.2.1 HIK-based Modified TFK

Let $\mathbf{y}_{in} = [y_{in1}, y_{in2}, \ldots, y_{inK}]^\top$ and $\mathbf{y}_{jm} = [y_{jm1}, y_{jm2}, \ldots, y_{jmK}]^\top$ be the nth and mth elements of the sequence of BOVW vectors $\mathbf{Y}_i$ and $\mathbf{Y}_j$ corresponding to the video sequences $\mathbf{X}_i$ and $\mathbf{X}_j$ respectively. The number of matches in the kth bin of the histogram is given by the histogram intersection function as

$$s_k = min(y_{ink}, y_{jmk}) \quad (2)$$

The HIK is computed as the total number of matches given by (Jan C. van Gemert and Zisserman, 2010)

$$K_{\text{HIK}}(\mathbf{y}_{in}, \mathbf{y}_{jm}) = \sum_{k=1}^{K} s_k \quad (3)$$

The HIK-based modified TFK is given by

$$K_{\text{HIKMTFK}}(\mathbf{Y}_i, \mathbf{Y}_j) = \sum_{n=1}^{N} \sum_{m=1}^{M} w_{nm} K_{\text{HIK}}(\mathbf{y}_{in}, \mathbf{y}_{jm}) \quad (4)$$

3.2.2 HK-based Modified TFK

In Hellinger's kernel the number of matches in the kth bin of the histogram is given by

$$s_k = \sqrt{y_{ink} y_{jmk}} \quad (5)$$

The Hellinger's kernel is computed as the total number of matches across the histogram and is given by

$$K_{\text{HK}}(\mathbf{y}_{in}, \mathbf{y}_{jm}) = \sum_{k=1}^{K} s_k \quad (6)$$

The HK-based modified TFK is given by

$$K_{\text{HKMTFK}}(\mathbf{Y}_i, \mathbf{Y}_j) = \sum_{n=1}^{N} \sum_{m=1}^{M} w_{nm} K_{\text{HK}}(\mathbf{y}_{in}, \mathbf{y}_{jm}) \quad (7)$$

3.3 Combining Kernels

To take the maximum advantage of the video representation, we consider the BOVW encoding of the

entire video $\mathbf{z}_i$ and the sequence of BOVW representation, $\mathbf{Y}_i$ of a video sequence, $\mathbf{X}_i$. We consider linear combination of kernels, MTFK and kernel computed on BOVW encoding of entire video.

$$K_{\text{COMB}}(\mathbf{X}_i, \mathbf{X}_j) = K_1(\mathbf{Y}_i, \mathbf{Y}_j) + K_2(\mathbf{z}_i, \mathbf{z}_j) \quad (8)$$

Here $K_1(\mathbf{Y}_i, \mathbf{Y}_j)$ is HIK-based MTFK or HK-based MTFK and $K_2(\mathbf{z}_i, \mathbf{z}_j)$ is LK or HK or HIK between BOVW vectors obtained from encoding whole video.

The base kernel (LK or HIK or HK) is a valid positive semidefinite kernel and multiplying a valid positive semidefinite kernel by a scalar is a valid positive semidefinite kernel (Shawe-Taylor, J. and Cristianini, N., 2004). Also the sum of valid positive semidefinite kernels is a valid positive semidefinite kernel (Shawe-Taylor, J. and Cristianini, N., 2004). The TFK and modified TFK both are valid positive semidefinite kernel.

4 EXPERIMENTAL STUDIES

In this section, we present the results of the experimental studies carried out to verify the effectiveness of the proposed kernels. The effectiveness of the proposed kernels are studied using UCF Sports dataset (Mikel D. Rodriguez and Shah, 2008; Soomro and Zamir, 2014) and UCF50 (Reddy and Shah, 2013) datasets. The UCF Sports dataset comprises of a collection of 150 sports videos of 10 activity classes. On an average, each class contains about 15 videos with an average length of each video being approximately 6.4 seconds. Since the dataset is small we follow the leave one out of strategy for activity recognition in videos as done in (Soomro and Zamir, 2014) and the video activity recognition accuracy presented corresponds to the average accuracy. The UCF50 dataset is obtained from YouTube videos. It comprises of 6681 video clips of 50 different activities (Mario Rodriguez and Makris, 2016). We follow the leave one out of strategy for activity recognition in videos this dataset and video activity recognition accuracy presented corresponds to the average accuracy.

Each video is represented using improved dense trajectories (IDT) descriptor (Wang and Schmid, 2013). The IDT descriptor densely samples feature points in each frame and tracks them in the video based on optical flow. To incorporate the temporal information, the IDT descriptor is extracted using a sliding window of 30 frames with an overlap of 15 frames. For a particular sliding window, multiple IDT descriptors each of 426 dimensions are extracted. The 426 features of an IDT descriptor comprise of multiple descriptors such as histogram of oriented gradient (HOG), histogram of optical flow (HOF) (Dalal and Triggs, 2005), and motion boundary histograms (MBH) (Dalal et al., 2006). The number of descriptors per window depends on the number of feature points tracked in that window. For video encoding, we propose to consider the GMM-based soft clustering approach. We also compare the GMM-based encoding approach with the codebook-based encoding approach proposed in (Mario Rodriguez and Makris, 2016). Different values for the codebook size in codebook-based encoding and the number of clusters in GMM, K was explored and an optimal value of 256 is chosen. In this work, we study the effectiveness of the proposed kernels for activity recognition by building SVM-based classifiers.

4.1 Studies on Activity Recognition in Video using BOVW Representation Corresponding to Entire Videos

In this section, we study the SVM-based activity recognition by encoding entire video into a single BOVW vector representation. For SVM-based classifier, we need a suitable kernel. We propose to consider linear kernel (LK), histogram intersection kernel (HIK) and Hellinger's kernel (HK). The performance of activity recognition in videos using SVM-based classifier is given in Table 1. The best performances are shown in boldface. It is seen from Table 1 that video activity recognition using SVM-based classifier using the frequency based kernels, HIK and HK is better than that using LK. This shows the suitability of frequency based kernels when the videos are encoded into BOVW representation. It is also seen that the performance of SVM-based classifier using the GMM-based encoding is better than that using the codebook-based encoding used in (Mario Rodriguez and Makris, 2016). This shows the effectiveness of GMM-based soft clustering approach for video encoding.

Table 1: Accuracy in (%) of SVM-based classifier for activity recognition in videos using linear kernel (LK), Hellinger's kernel (HK) and histogram intersection kernel (HIK) on the BOVW encoding of the entire video. Here CBE corresponds to code book based encoding proposed in (Mario Rodriguez and Makris, 2016) and GMME corresponds to GMM-based video encoding proposed in this work. The results presented correspond to the BOVW encoding for the entire video.

Kernel	UCF Sports		UCF50	
	CBE	GMME	CBE	GMME
LK	80.67	90.40	80.40	90.40
HK	83.33	**92.93**	**83.33**	92.27
HIK	**84.67**	92.53	82.27	**92.53**

4.2 Studies on Activity Recognition in Videos using Sequence of BOVW Vectors Representation for Videos

In this section, we study the activity recognition in videos using sequence of BOVW vectors representation of videos. Each video clip is split into a set of segments. Each segment is encoded into a BOVW vector using the codebook based encoding (Mario Rodriguez and Makris, 2016) and GMM-based encoding. For SVM-based classifier, we consider time flexible kernel (TFK) and modified TFKs (MTFKs) using the frequency based kernels HK and HIK for activity recognition. The performance of SVM-based activity recognizer using TFK and MTFKs is given in Table 2. The best performance is shown in boldface. It is seen from Table 2 that MTFK-based SVMs give better performance than TFK-based SVMs. This shows the effectiveness of the kernels proposed in this work. From Tables 1 and 2, it is seen that the TFK-based SVMs give better performance than LK-based SVMs that uses BOVW vector representation of entire video. This shows the importance of using the temporal information of video for the activity recognition. It is also seen that performance of MTFK-based SVMs is better than the SVM-based classifiers using the frequency based kernels, HK and HIK on BOVW vectors encoding corresponding to the entire video.

Table 2: Accuracy in (%)of SVM-based classifier for activity recognition in videos using time flexible kernel (TFK) and modified TFKs (MTFKs) computed on sequence of BOVW vectors representation of videos. Here CBE corresponds to code book based encoding proposed in (Mario Rodriguez and Makris, 2016) and GMME corresponds to GMM-based video encoding proposed in this work. HKMTFK corresponds to HK-based modified TFK and HIKMTFK denotes the HIK-based modified TFK.

Kernel	UCF Sports		UCF50	
	CBE	GMME	CBE	GMME
TFK	82.00	91.27	81.67	91.27
HKMTFK	**86.67**	**95.73**	**86.67**	**95.27**
HIKMTFK	86.00	95.60	86.00	95.00

4.3 Studies on Activity Recognition in Video using Combination of Kernels

In this section, we combine a kernel computed on BOVW representation of entire videos and a kernel computed on sequence of BOVW vectors representation of video. We consider simple additive combination so that a combined kernel COMB($K_1 + K_2$) corresponds to addition of K_1 and K_2. Here K_1 corresponds to kernel computed using sequence of BOVW vectors representation of videos, and K_2 corresponds to the kernel computed on the BOVW representation of entire video. The performance of SVM-based classifier using combined kernels for video action recognition is given in Table 3. The best performances are given in boldface. It is seen from Tables 2 and 3 that the accuracy of SVM-based classifier using the combination kernel involving TFK is better than that for the SVM-based classifier using only TFK. This shows the effectiveness of the combination of kernels. It is also seen that the performance for SVM-based classifier using combination of kernels involving HK and HIK computed on entire video is better than that obtained with combination of kernel involving LK computed on entire video. It is also seen that the performance of SVM-based classifiers using combination kernel involving MTFKs is better than that using TFKs. This shows the effectiveness of the proposed MTFK in activity recognition in videos.

Table 3: Comparison of performance of activity recognition in video using SVM-based classifiers using combination of kernels on UCF sports dataset. Here, COMB($K_1 + K_2$) indicate additive combination of kernels K_1 and K_2 respectively. K_1 is a kernel computed on the sequence of BOVW vectors representation of videos and K_2 is a kernel computed on the BOVW representation of the entire video. Here CBE corresponds to code book based encoding proposed in (Mario Rodriguez and Makris, 2016) and GMME corresponds to GMM-based video encoding proposed in this work.

Kernel	UCF Sports		UCF50	
	CBE	GMME	CBE	GMME
COMB(TFK+LK)	82.67	93.87	82.00	93.87
COMB(HKMTFK+LK)	84.67	94.13	84.67	93.87
COMB(HIKMTFK+LK)	82.67	94.13	83.67	93.87
COMB(TFK+HK)	86.00	94.13	85.13	94.13
COMB(HKMTFK+HK)	86.67	96.53	86.00	96.13
COMB(HIKMTFK+HK)	**87.33**	**96.93**	**86.93**	96.27
COMB(TFK+HIK)	84.00	94.67	84.00	94.67
COMB(HKMTFK+HIK)	86.67	96.40	86.67	**96.40**
COMB(HIKMTFK+HIK)	86.00	**96.93**	86.27	96.27

Next we compare the SVM-based classification performance using the various kernels used in this study. The performance of SVM-based classifiers using different kernels for video activity recognition in UCF Sports dataset is given in Figure 1. The performance of SVM-based classifiers using different kernels for video activity recognition in UCF50 dataset is given in Figure 2. Here, we consider on GMM-based video encoding since it is seen from Tables 1, 2 and 3 that GMM-based encoding is better than codebook based encoding proposed in (Mario Rodriguez and Makris, 2016). It is seen from Figures 1 and 2 the

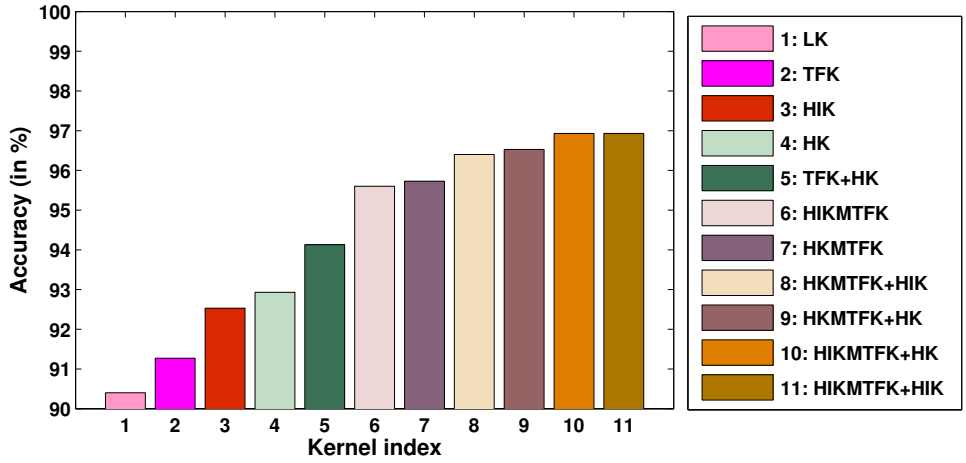

Figure 1: Comparison of performance of SVM-based classifier using different kernels for the video activity recognition in UCF Sports dataset.

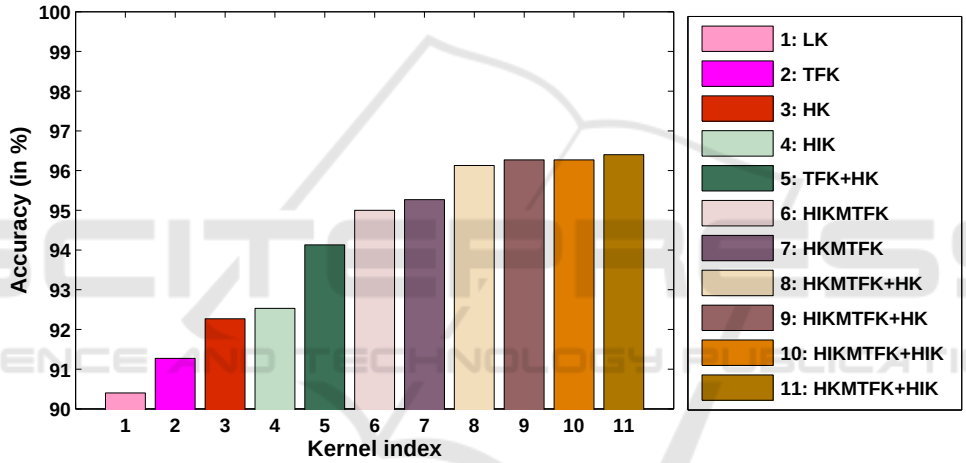

Figure 2: Comparison of performance of SVM-based classifier using different kernels for the video activity recognition in UCF50 dataset.

proposed modified TFKs are effective for the video activity recognition.

5 CONCLUSIONS

In this paper, we proposed modified time flexible kernel (MTFK) for SVM-based video activity recognition. We also proposed a GMM-based video encoding approach to encode a video into a sequence of bag of visual words (BOVW) vectors representation. The computation of MTFK between two videos involves matching every pair of BOVW vectors corresponding to the videos. In this work, we explored the use of frequency based kernels such as histogram intersection kernel (HIK) and Helliger's kernel (HK) to match a pair of BOVW vectors. The studies conducted on benchmark datasets show that MTFKs using HK and HIK are effective for the activity recognition in videos. In this work, we explored representing a video using improved dense trajectories which is a low-level spatio temporal video representation. In future, the proposed MTFK need to be studied using the other approaches for video representation such as deep learning based representation etc.

REFERENCES

Adrien Gaidon, Z. H. and Schmid, C. (2012). Recognizing activities with cluster-trees of tracklets. In *Proceedings of British Machine Vision Conference (BMVC 2012)*, pages 30.1–30.13.

Alexander Klaser, M. M. and Schmid, C. (2008). A spatio-temporal descriptor based on 3D-gradients. In

Proceedings of British Machine Vision Conference (BMVC 2008), pages 995–1004.

Chatfield Ken, Lempitsky Victor S, V. A. and Andrew, Z. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, volume 2, page 8.

Cuiwei Liu, X. W. and Jia, Y. (2016). Transfer latent svm for joint recognition and localization of actions in videos. *IEEE transactions on cybernetics*, 46(11):2596–2608.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference*, volume 1, pages 886–893. IEEE.

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

Dan Oneata, J. V. and Schmid, C. (2013). Action and event recognition with Fisher vectors on a compact feature set. In *Proceedings of IEEE International Conference on Computer Vision (ICCV 2013)*.

Dan Xu, Xiao Xiao, X. W. and Wang, J. (2016). Human action recognition based on kinect and pso-svm by representing 3d skeletons as points in lie group. In *International Conference on Audio, Language and Image Processing (ICALIP), 2016 International Conference on*, pages 568–573. IEEE.

Dileep, A. D. and Chandra Sekhar, C. (2013). Hmm based intermediate matching kernel for classification of sequential patterns of speech using support vector machines. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2570–2582.

Jagadeesh, B. and Patil, C. M. (2016). Video based action detection and recognition human using optical flow and svm classifier. In *Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference on*, pages 1761–1765. IEEE.

Jan C. van Gemert, Victor S. Lempitsky, A. V. and Zisserman, A. (2010). Visual word ambiguity. *IEEE transactions on pattern analysis and machine intelligence*, 32(17):1271–1283.

Jin Wang, Ping Liu, M. F. S. and Liu, H. (2011). Hman action categoriztaion using condition random fileds. In *Proceedings of Robotic Intelligence in Informationally Structures Space (RiiSS)*.

Junji Yamato, J. O. and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 1992)*, pages 1063–1069.

Laptev, I. (2005). On space-time interest points. *International Journal on Computer Vision*, 64:107–123.

Mario Rodriguez, Orrite Carlos, C. M. and Makris, D. (2016). A time flexible kernel framework for video-based activity recognition. *Image and Vision Computing*, 48:26–36.

Mikel D. Rodriguez, J. A. and Shah, M. (2008). Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE COnference on Computer Vision and Pattern Recogniton (CVPR 2008)*, pages 1–8. IEEE.

Neeraj Sharma, Anshu Sharma, V. T. and A.D, D. (2016). Text classification using combined sparse representation classifiers and support vector machines. In *Proceedings of the 4th International Symposium on COmputational and Business Intelligence*, pages 181–185, Olten, Switzerland.

Orit Kliper-Gross, Yaron Gurovich, T. H. and Wolf, L. (2012). Motion interchange patterns for action recognition in unconstrained videos. In *European Conference on Computer Vision*, pages 256–269. Springer.

Paul Scovanner, S. A. and Shah, M. (2007). A 3-dimensional SIFT descriptor and its application to action recognition. In *Proceedings of the 15th ACM International Conference on Multimedia*, pages 357–360. ACM.

Reddy, K. K. and Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision Applications*, 24:971–981.

Shabou, A. and LeBorgne, H. (2012). Locality-constrained and spatially regularized coding for scene categorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference*, pages 3618–3625. IEEE.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K.

Soomro, K. and Zamir, A. R. (2014). Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558.

Zhou, Y. and Fan, C. (2016). Action recognition robust to position changes using skeleton information and svm. In *International Conference on Robotics and Automation Engineering (ICRAE),*, pages 65–69. IEEE.