

# Spatial information in classification of activity videos

Shreeya Sengupta\*<sup>†</sup>, Hui Wang\*<sup>‡</sup>, William Blackburn\*<sup>§</sup> and, Piyush Ojha\*<sup>¶</sup>

\*School of Computing and Mathematics, University Of Ulster,

Northern Ireland, UK, BT370QB

<sup>†</sup> sengupta-s@email.ulster.ac.uk

<sup>‡</sup> h.wang@ulster.ac.uk

<sup>§</sup> wt.blackburn@ulster.ac.uk

¶ pc.ojha@ulster.ac.uk

Abstract-Spatial information describes the relative spatial position of an object in a video. Such information may aid several video analysis tasks such as object, scene, event and activity recognition. This paper studies the effect of spatial information on video activity recognition. The paper firstly performs activity recognition on KTH and Weizmann videos using Hidden Markov Model and k-Nearest Neighbour classifiers trained on Histogram Of Oriented Optical Flows feature. Histogram of Oriented Optical Flows feature is based on optical flow vectors and ignores any spatial information present in a video. Further, in this paper, a new feature set, referred to as Regional Motion Vectors is proposed. This feature like Histogram of Oriented Optical Flow is derived from optical flow vectors; however, unlike Histogram of Oriented Optical Flows preserves any spatial information in a video. Activity recognition was again performed using the two classifiers, this time trained on Regional Motion Vectors feature. Results show that when Regional Motion Vectors is used as the feature set on the KTH dataset, there is a significant improvement in the performance of k-Nearest Neighbour. When Regional Motion Vector is used on the Weizmann dataset, performances of the k-Nearest Neighbour improves significantly for some of the cases and for the other cases, the performance is comparable to when oriented optical flows is used as a feature set. Slight improvement is achieved by Hidden Markov Model on both the datasets. As Histogram of Oriented Optical Flows ignores spatial information and Regional Motion Vectors preserves it, the increase in the performance of the classifiers on using Reginal Motion Vectors instead of Histogram of Oriented Optical Flows illustrates the importance of spatial information in video activity recognition.

# I. INTRODUCTION

**S** PATIAL information describes the physical position of an object and its spatial relationship to other objects. It plays a crucial role in video activity recognition and may be very useful in differentiating between static and mobile activities. Mobile activities are activities where a person performing the activity moves along the field of view and static activities are activities where the person performing the activity remains at one place. Spatial information provides the position of a person and thus considering spatial information while activity recognition is expected to aid differentiation between static and mobile activities.

The importance of spatial information has also been studied by a group of researcher in Amsterdam [1]. According to them, the spatial extent of an object depends on the object to be classified itself. Spatial extent captures contextual information and for some objects the spatial extent is the whole scene whereas for some the extent is to a specific rigid boundary. For identifying functionalities such as walking or jumping, the more the context, the better it is. For identifying objects such as a car, a plane or a bus, the context is the object only.

In video activity recognition literature spatial information is often captured by various local space-time features as defined in [2], [3], [4], [5], [6], [7], [8], [9], [10], [11] and [12]. These local space-time features capture frame-wise spatial information by first detecting interest points with either interest point detectors (Harris detector, Hessian detectors, edge detector, corner detectors) or various sampling methods (dense sampling [13] or motion adaptive sampling [14]) for each frame, then spatio-temporal regions are defined around all the detected points in each frame and finally the spatio-temporal regions are described using one of the local space-time features. Other attempts to capture spatial information is contextual bag of words (CBOW) [15] and BOW with spatial pyramid [16]. Both [15] and [16] are extensions of the bag of words(BOW) feature set. BOW is a frequency based feature set which was initially used for text classification where it represents the occurrence of words in a text document. It has now been adapted in computer vision where it represents a video by counting the occurrence of a visual word in the video. Thus, BOW is a frequency based descriptor and it ignores any spatial relationship between the visual words. The extensions [15] and [16] were proposed to incorporate the missing spatial relationship explicitly in BOW.

Following the local space-time approaches this paper proposes a new feature set to preserve spatial information in activity recognition data. The new feature set, referred to as regional motion vectors (RMV), is based on optical flow vectors. Evaluation of the new set on KTH videos shows significant improvement in classification accuracy when compared with histogram of oriented optical flows (HOOF), an existing optical flow based feature set which does not preserve spatial information. The new feature set has also been assessed on the Weizmann dataset.

The next section (Section II) explains the proposed methodology, followed by the similar work section (Section III), the experimental setup section (Section IV), the results section (Section V), the conclusion (Section VI) section and finally the future works section (Section VII).

# II. PROPOSED METHODOLOGY

In this paper a new feature set based on optical flow vectors has been proposed. The new feature set is derived such that it preserves the spatial information present in a video. RMV was derived by dividing a frame f of resolution  $n \times m$  into subregions using a grid of resolution  $r \times s$ , adding the magnitude of the optical flow vectors in each sub-region, and normalising the sum of RMVs to unity. Thus the overall relative motion  $\psi$ of a sub-region  $SR_{(a,b)}$  in a frame was computed as shown in Equation 1

$$\psi(SR_{(a,b)}) = \sum_{i=(a-1)r+1}^{ar} \sum_{j=(b-1)s+1}^{bs} |OFV(i,j)|/N, \quad (1)$$

where,

 $a \in \{1, 2, \cdots, n_r\},\ b \in \{1, 2, \cdots, m_s\},\$ 

 $n_r = \lfloor \frac{n}{r} \rfloor,$   $m_s = \lfloor \frac{m}{s} \rfloor,$  OFV(i, j): Optical Flow Vector of the  $i^{th}$  row and  $j^{th}$ column.

 $SR_{(a,b)}$ : a sub-region  $SR_{(a,b)}$ ,

N: normalisation constant.

Equation 1 was applied to all the sub-regions in a frame, thus producing a column vector  $\Psi_t$  as shown in Equation 2:

$$\Psi_t = \{\psi(SR_{(a,b)})\}, a \in \{1, 2, \cdots, n_r\}, b \in \{1, 2, \cdots, m_s\},$$
(2)

where,

 $\psi(SR_{(a,b)})$ : the motion of sub-region  $SR_{(a,b)}$ .

All such  $\Psi_t$  were concatenated to form the RMV feature set (Equation 3):

$$RMV = \{\Psi_1, \Psi_2, \cdots, \Psi_T\}.$$
(3)

In this method, spatial information was preserved in the videos by dividing a frame into several sub-regions  $SR_{(a,b)}$ ,  $a \in \{1, 2, \cdots, n_r\}$  and  $b \in \{1, 2, \cdots, m_s\}$ . As the subregions were spatially correlated, any information extracted from these regions inherited the spatial relationship from the regions. Thus, the spatial information was not lost.

The next section (Section III) describes some of the existing features that are similar to the feature proposed in this paper.

## III. SIMILAR WORK

Histogram of oriented optical flow (HOOF) proposed by Chaudhary et al. in [4] is a widely used feature set for video activity recognition. It is similar to RMV because both of them are calculated from raw optical flow vectors; however, unlike RMV, HOOF does not divide a frame into sub-regions. To extract HOOF, firstly optical flow vectors are calculated using either Horn-Schunck or Lucas-Kanade algorithm, then the flow vectors are binned into ninety angular bins, ranging from  $-180^{\circ}$  to  $+180^{\circ}$ , according to their orientation and finally, the magnitude of the vectors in each of the bins is summed. Thus, while HOOF is a measure of motion in some specified directions (defined by the angular bin range) in each frame. RMV is a relative measure of motion of each subregion in a frame. Also, while any spatial correlation among the flow vectors is lost in HOOF due to the binning strategy which ignores the spatial positioning of the flow vectors, RMV preserves such correlation by dividing a frame into several spatially correlated sub-regions.

Another feature set for representing a video which is very similar to the feature set proposed in this paper is the feature proposed by Janez Pers et al. in [11]. Similar to RMV proposed here, derivation of their representation also included dividing a frame of a video into various sub-regions. However, after the division, they calculated HOOF in each of the subregions unlike ours where only the relative motion of each of the sub-regions was calculated. Further, they converted the calculated HOOFs into a sequence of symbols and their final representation of a video was a sequence of symbols, the final representation of our method was a sequence of relative motions of various sub-regions in each frame of a video. Perz et al. proposed a frequency based representation of videos whereas this paper proposes a motion based representation of videos.

Raw optical flow vectors have also been used to derive space time appearance (STA) descriptor proposed in [17]. The computation of STA descriptors in [17] commenced by detecting interest regions in a video and then, the detected regions were divided into sub-regions. Dense optical flow vectors were calculated using the Farneback and TV-L1 optical flow algorithms and grid histograms representing the distribution of the optical flow vectors were computed in each sub-region. Grid histograms were concatenated to form the grid vectors and a weighted average of the grid vectors formed the order one STA (STA1) descriptors. Order two STA (STA2) descriptors were then obtained by combining the grid histograms to form component vectors and then binning the component vectors into k2 bins. The final feature was obtained by concatenating the STA2 descriptors into a vector. The only similarities between STA and RMV are the division of interest regions into sub-regions and use of optical flow vectors. While STA is again a frequency based approach which represents the distribution of optical flow vectors upto two orders, RMV does not represent any such distribution and only measures the motion of sub-regions in a frame.

While the features proposed in [11] and [17] preserve spatial information, HOOF in [4] does not preserve any spatial information. As the main aim of the proposed method was to preserve the spatial information in video data to aid activity recognition, the effectiveness of the proposed method has been studied by comparing its performance only with HOOF. HOOF is a feature set derived from optical flow vectors without any spatial information and the proposed method (RMV) is a feature set again derived from optical flow vectors but with spatial information. Thus, a comparison between the two methods is expected to illustrate the effectiveness of the proposed method as well as the importance of spatial information in video activity recognition.

The next section (Section IV) describes the experimental setup for comparing both the methods.

## IV. EXPERIMENTAL SETUP

RMV and HOOF were tested on videos from the KTH dataset [18] and the Weizmann dataset [19].

The KTH dataset is a video dataset consisting of six different human actions, namely boxing, hand-waving, hand-clapping, jogging, running and walking. These six actions were performed by twenty five subjects under four different scenarios: outdoors, outdoors with varying scale, outdoors with subjects wearing a variety of clothes and indoors. All the videos were taken over homogenous backgrounds with a static camera and a frame rate of twenty five frames per second. For this study, each video of this dataset was further divided into four sub-videos, and therefore, with twenty five people, six actions, four scenarios and four sub-videos, there are in total 2400 sub-videos in the dataset. Out of these, 120 sub-videos of each action were selected randomly, thus providing a total of 720 sub-videos for experimentation.

The Weizmann dataset [19] contains nine people performing ten different actions: gallop, jump, walk, run, gallop sideways, bend, one hand waving, jumping jack,two hand waving, jumping in place and skip. The actions were recorded at a resolution of  $180 \times 144$ .

For both the datasets in this study, the optical flow vectors for deriving HOOF were obtained using the Lucas-Kanade algorithm [20]. The optical flow vectors were then sorted into ninety angular bins, each 4° wide, collectively covering the full angular range from  $-180^{\circ}$  to  $180^{\circ}$ . The magnitudes of the optical flow vectors in each bin were added to produce a ninety dimensional optical flow vector (or histogram) for each frame. Thus for a *T* frame sequence, we get a  $90 \times T$ dimensional matrix referred to as HOOF. Principle component analysis was then used to reduce the data dimension to five, eight, twelve and sixteen.

For deriving regional motion vectors (RMV) feature set, the Lucas-Kanande algorithm was used to compute frame-wise optical flow vectors where each vector again had two dimensions - the magnitude and direction. Then, for the KTH dataset, instead of the angular bins, each frame was divided into subregions  $SR_{(a,b)}, a \in \{1, 2, \dots, n_r\}$  and  $b \in \{1, 2, \dots, m_s\}$ by using a patch of resolution r = 10 by s = 20. As the resolution of each frame was n = 120 by m = 160, so, each frame was divided into (120/10) \* (160/20) = 96 regions. The vectors in each of these regions were grouped into one bin and their magnitudes were added. Vector sum of these vectors could also have been considered. However, as there was no significant difference, only the results obtained using magnitude have been listed. The value of r and s (size of the patch) could also have been varied. However, the size of the patch was chosen such that the number of bins was near to ninety - the number of bins in HOOF. To summarize, RMV produced a  $96 \times 1$  dimensional column vector for each frame, and for a sequence of T frames, a  $96 \times T$  dimensional matrix. This matrix was known as regional motion vectors (RMV). Principle component analysis was again used to reduce the data dimension to five, eight, twelve and sixteen.

Similarly, RMV features for the Weizmann action videos were obtained by dividing each frame into sub-regions instead of angular bins. Again, a patch was used for the purpose, however, with a different resolution. The resolution of the patch used was r = 18 by s = 16 and since, the resolution of the frames were n = 180 by m = 144,  $90 \times 1$  dimensional column vector was produced for each frame. For a sequence of T frames,  $90 \times T$  dimensional RMV matrix was produced. The dimension of the matrix was reduced to five, eight, twelve and sixteen using principal component analysis.

Once HOOF and RMV features were extracted, they were used with k-nearest neighbour (kNN) and hidden Markov model (HMM) classifiers. An unclassified pattern is assigned to the class of its nearest neighbour. The similarity between two points in the multi-dimensional space was defined either via their Euclidean distance (EUC) or by a neighbourhood counting similarity metric (NCM) [21]. These two measures can be extended directly to patterns, i.e. sequences of points. (When computing the Euclidean distance between sequences of unequal length, we truncate the longer sequence so that its length matches the shorter sequence.) Alternatively, EUCand NCM can be used as the underlying point-to-point similarity measures in dynamic time warping (DTW) [22], longest common subsequence (LCS) [23] or all common subsequences (ACS) [24] measures of similarity between sequences. Thus, eight measures of similarity were evaluated between sequences: EUC, NCM, DTW + EUC, DTW +NCM, LCS + EUC, LCS + NCM, ACS + EUC and ACS + NCM. The performances of these measures were evaluated because some of these measures are widely used and are known to handle variation in time series data well. The kNN classifier along with the similarity measures were coded from scratch.

HMM was used with three Gaussian and six states and EUC was the default similarity measure for this model. HMM was implemented using the Kevin Murphy toolbox [25] for HMM.

The classification regime was ten fold cross validation. The dataset was arranged such that the test set contained one video from each action category and the remaining videos from those categories were used as the training set. The following section (Section V) lists the performance of the two classifiers.

## V. RESULTS

This section lists the performance of all the classifiers using HOOF and RMV feature set on KTH and Weizmann videos. In this study HOOF is the optical flow feature set which lacks spatial information, i.e. any spatial relationship in the videos is lost when HOOF is extracted. On the contrary, when RMV feature is extracted, spatial correlations in the videos are also preserved. Thus, it is expected that a comparison of the performances of several classifiers(HMM and kNN in this study) using HOOF and RMV will illustrate the importance of considering spatial information during video activity recognition.

In this section, first the performance of the classifiers on the KTH dataset is listed, followed by the performance of the classifiers on the Weizmann dataset.

#### A. Performance of HMM and kNN on KTH dataset

Table I presents the performance of kNN and HMM on the KTH dataset. In the table the first column represents the feature set used (RMV or HOOF) for activity recognition, the second column lists the classifiers used. It can be noted that several similarity measures (EUC, NCM, DTW + EUC, DTW + NCM, LCS + EUC, LCS + NCM, ACS + EUC and ACS + NCM) have been explicitly specified with kNN in the second column. The measures indicate the method used with kNN to calculate similarity between two sequences while recognising activities. HMM was used with only EUC distance measure and hence the measure has not been specified explicitly. The following columns three to six list the performance of both the classifiers on varying dimension of HOOF and RMV data. The varying dimensions have been indicated in the header as 5 PCs, 8 PCs, 12 PCs and 16 PCs. Here, PC stands for principle components and 5, 8, 12 and 16 stands for the number of principle components selected.

TABLE I. THE PERFORMANCE OF HMM & KNN WITH VARIOUS SIMILARITY MEASURES ON KTH DATASET USING BOTH THE HOOF AND THE RMV FEATURES. IN THE COLUMN HEADINGS, PC STANDS FOR PRINCIPLE COMPONENTS.

Feature set	Classifiers	5PCs	8PCs	12PCs	16PCs
	НММ	59	63	68	68
HOOF	kNN + EUC	59	64	67	66
	kNN + NCM	19	17	16	16
	kNN + (DTW + EUC)	60	67	70	72
	kNN + (DTW + NCM)	21	22	24	25
	kNN + (LCS + EUC)	26	25	25	24
	kNN + (LCS + NCM)	23	26	25	23
	kNN + (ACS + EUC)	22	21	21	21
	kNN + (ACS + NCM)	21	21	22	19
RMV	НММ	61	67	72	71
	kNN + EUC	86	88	89	90
	kNN + NCM	70	56	50	43
	kNN + (DTW + EUC)	77	77	78	79
	kNN + (DTW + NCM)	33	38	45	50
	kNN + (LCS + EUC)	34	38	33	16
	kNN + (LCS + NCM)	34	37	40	46
	kNN + (ACS + EUC)	33	35	33	33
	kNN + (ACS + NCM)	33	34	33	33

Following Table I, the performance of HMM and kNN on varying dimensional HOOF has been compared with their performance of varying dimensional RMV. The comparisons have been represented graphically in Figures 1, 2, 3 and 4. In the figures, classifiers are listed on the x-axis and the classification accuracy (activity recognition rate) achieved by them is listed on the y-axis.

Figure 1 compares the performance of HMM and kNN classifiers using five dimensional HOOF with their performance using five dimensional RMV data. It can be observed that performance of both HMM and KNN improves when RMV is used instead of HOOF. For HMM the improvement is marginal from 59% when HOOF is used to 61% when RMV is used. Significant improvement is noticed in cases where kNN was used with *EUC* and *NCM* similarity measures. In case of kNN with *EUC* the change was from 59% to 86% and in case of kNN with *NCM* it was from 18% to 70%. The improvement in accuracy when RMV is used instead of HOOF indicates the importance of preserving spatial information in video data.

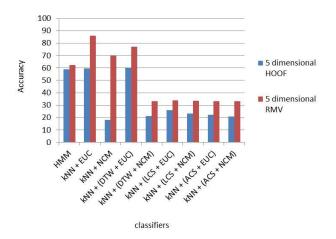


Fig. 1. Comparison of the performances of different classifiers on five dimensional HOOF and RMV data.

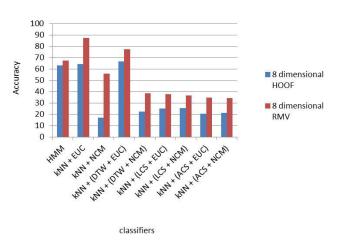


Fig. 2. Comparison of the performances of different classifiers on eight dimensional HOOF and RMV data.

Figure 2 presents a comparison similar to Figure 1. How-

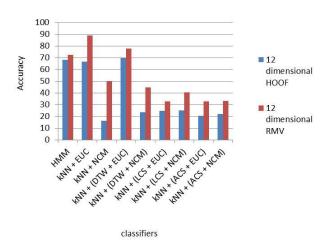


Fig. 3. Comparison of the performances of different classifiers on twelve dimensional HOOF and RMV data.

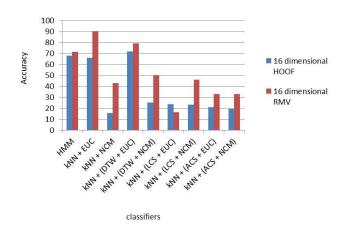


Fig. 4. Comparison of the performances of different classifiers on sixteen dimensional HOOF and RMV data.

ever, instead of five dimensional HOOF and RMV data, eight dimensional HOOF and RMV was used. It is observed that again an improvement in classification accuracy is achieved when RMV is used instead of HOOF. When RMV is used with kNN for activity recognition EUC and NCM similarity measures perform significantly better than when HOOF is used with kNN. kNN + EUC achieved an accuracy of 64% when HOOF is used (spatial information in videos is ignored) and it achieves an accuracy of 88% when RMV is used (spatial information in KTH videos is considered). The accuracy obtained by kNN + NCM in absence of spatial information (HOOF feature set is used) is 17% and in presence of spatial information (RMV feature is used) is 56%. The results thus again show the importance of considering such spatial information.

Trend similar to Figure 1 and Figure 2 is also noted in Figure 3 and Figure 4, illustrating the importance of considering spatial information in videos.

It is also noted that in Figures 1, 2, 3 and 4 kNN with DTW + EUC performs well consistently. When HOOF is used the accuracies are 60%, 67%, 70% and 72% for five, eight, twelve and sixteen dimensional HOOF data and when RMV is used the accuracies are 77%, 77%, 78% and 79% again for five, eight, twelve and sixteen dimensional data. Thus, for DTW + EUC the accuracies are relatively stable through varying dimensional data. This was expected because DTW is designed in such a way that it calculates a similarity score between two given sequences by matching each element of one sequence with every element of the other sequence. Such a matching facilitates comparing sequences having different number of frames which is very common in video data. Varying number of frames introduce a different type of variation which also hinders activity recognition. As DTW handles such variation, its performance is relatively steady and superior to other similarity measures (for example when the data is five dimensional data, DTW + EUC with kNN gives an accuracy of 60%which is higher than other measures such as NCM (18%), LCS + EUC (26%), LCS + NCM (23%), ACS + EUC(22%) and ACS + NCM(21%).) during activity recognition.

However, it is also observed that although DTW + EUCoutperformed all the measures but DTW + NCM performed poorly. This can be attributed to the different underlying point to point similarity measure that has been used with DTW. When the performance of EUC and NCM alone is compared, it is observed that when HOOF is used EUCperforms significantly better than NCM with an accuracy of 59% (accuracy obtained using NCM is 18%). A possible explanation for such a performance of NCM is its ability to work better on correlated data than uncorrelated data. This explanation is supported by the improvement in NCM's performance when RMV, where the data is spatially correlated is used instead of HOOF, where the data is uncorrelated. The accuracy obtained by NCM when  $\ensuremath{\mathsf{RMV}}$  is used is 70% and when HOOF is used is 18%. The behaviour of NCM on correlated and uncorrelated data extends to DTW when these two measures are used as the underlying point to point measure with DTW. Thus, DTW + EUCperforms superior to DTW + NCM. This behaviour of NCM also supports considering spatial relationships and preserving spatial information during video activity recognition.

The next subsection lists and reviews the performance of kNN and HMM on Weizmann videos in the presence and absence of spatial information. Again, presence of spatial information is ensured by using RMV as the feature set and absence of the information is ensured by using HOOF.

#### B. Performance of kNN and HMM on Weizmann dataset

In this subsection table II shows the performance of kNN and HMM using HOOF and RMV features. Similar to table I, column one shows the feature set being used, column 2 lists the classifiers and the rest of the columns (3-6) lists the obtained classification accuracy with varying dimensional HOOF and RMV.

Following the table are figures 5, 6, 7 and 8 which compare the performances of the classifiers while using five, eight, twelve and sixteen dimensional HOOF and RMV respectively.

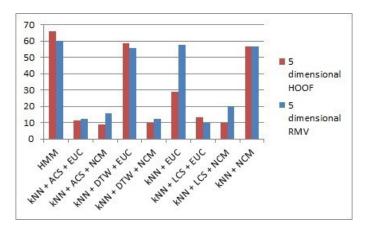


Fig. 5. Comparison of the performances of different classifiers on Weizmann dataset using five dimensional HOOF and RMV.

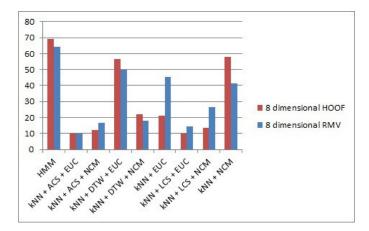


Fig. 6. Comparison of the performances of different classifiers on Weizmann dataset using eight dimensional HOOF and RMV.

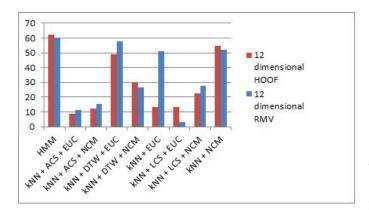


Fig. 7. Comparison of the performances of different classifiers on Weizmann dataset using twelve dimensional HOOF and RMV.

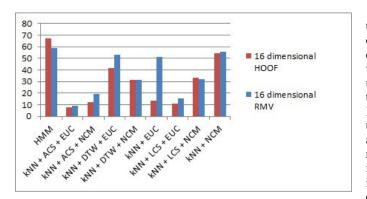


Fig. 8. Comparison of the performances of different classifiers on Weizmann dataset using sixteen dimensional HOOF and RMV.

TABLE II.	THE PERFORMANCE OF HMM & KNN WITH VARIOUS
SIMILARITY MEA	ASURES ON WEIZMANN DATASET USING BOTH THE HOOF
AND THE <b>RMV</b>	FEATURES. IN THE COLUMN HEADINGS, PC STANDS FOR
	PRINCIPLE COMPONENTS.

Feature set	Classifiers	5PCs	8PCs	12PCs	16PCs
	HMM	66	69	62	67
HOOF	kNN + EUC	29	21	13	13
	kNN + NCM	57	58	54	54
	kNN + (DTW + EUC)	59	57	49	41
	kNN + (DTW + NCM)	10	22	30	31
	kNN + (LCS + EUC)	13	10	13	11
	kNN + (LCS + NCM)	10	13	22	33
	kNN + (ACS + EUC)	11	10	9	8
	kNN + (ACS + NCM)	9	12	12	12
	НММ	60	64	60	59
RMV	kNN + EUC	58	46	51	51
	kNN + NCM	57	41	52	56
	kNN + (DTW + EUC)	56	50	58	53
	kNN + (DTW + NCM)	12	18	27	31
	kNN + (LCS + EUC)	10	14	3	16
	kNN + (LCS + NCM)	20	27	28	32
	kNN + (ACS + EUC)	12	10	11	9
	kNN + (ACS + NCM)	16	17	16	19

From the figures, it can be seen that on several occasions, the performance of kNN increases when RMV is used instead of HOOF. An example is the improvement in classification accuracy from 29 to 58 for kNN+EUC when 5PC RMV is used instead of 5PC HOOF. Another example is the performance of kNN + (DTW+EUC) using 12 dimensional features. When 12 dimensional HOOF is used, the classification accuracy is 49, and when 12 dimensional RMV is used, the classification accuracy is 58. However, for some cases, the performance of the classifiers are either comparable or remains the same. For example, the highest classification accuracy obtained using kNN+(DTW+EUC) and 5 dimensional HOOF feature set is 59. However, when 5 dimensional RMV is used for the same case, the accuracy is 56. Then the performance of kNN + NCM on 5 dimensional HOOF is 57 which remains unchanged when 5 dimensional RMV is used.

Therefore, from the above results it can be observed that there is a significant improvement in the performance of the classifiers when RMV is used instead of HOOF on the KTH dataset. However, no such significant performance difference was noted for the classifiers when tested on Weizmann dataset using RMV and HOOF features. This can be attributed to the lack of direction information in the RMV feature set. Incorporating direction information in the feature set ensures that the direction (left to right or right to left) of mobile actions (jog, run, skip, walk) does not affect the classification / recognition results. As there are less number of mobile actions in the KTH dataset than the Weizmann dataset, lack of this information does not affect the overall performance of the classifiers on the KTH dataset. However, a closer look at the confusion matrix of the KTH dataset reveals that most of the misclassification is among the mobile activities, which in case of RMV can be attributed to the absence of direction information in the RMV feature set.

The competitive and sometimes better performance of all the classifiers on the Weizmann dataset does not reduce the significance of the proposed feature set. On the contrary, it shows the potential of the feature by performing well (on KTH dataset) and at par (on Weizmann dataset), even after the lack of direction information. This shows that beside direction, spatial relationships also play an important role in action classification and, thus, features preserving such spatial relationships are required.

Finally, it can be observed that the dimension of the data has been varied from five to sixteen. Data higher than sixteen dimension was not considered because the performance of the classifiers for most of the cases attains stability after eight dimensional HOOF and RMV data. Further, the computation time of kNN increases with increasing dimension and hence considering very high dimensional data is undesirable. Last but not the least, training HMM on such high dimensional data is not only time consuming but requires large number of training samples. As huge number of data may not always be available, data with dimension higher than sixteen is not considered.

#### VI. CONCLUSION

The main aim of this paper was to study the effects of spatial information in video data analysis. For this, the paper focussed on video activity recognition and videos from the KTH and Weizmann datasets, which are activity datasets, were selected for this purpose. Recognition was performed using two classifiers - kNN and HMM, trained firstly on HOOF and then on RMV features. HOOF is an optical flow based feature set which ignores any spatial information and therefore applying any classifier on such a feature set illustrates a scenario where spatial information has been ignored (absence of spatial information). RMV on the contrary is an optical flow based feature set where spatial information is preserved. Thus, any spatial relationships among objects in a video or object and background in a video is present in the feature set. A comparison of the performances of the classifiers with HOOF and RMV illustrated the effect of the presence and absence of spatial information on video activity recognition. Results show significant improvement in the performance of kNN classifier on KTH dataset, when spatial information is preserved. For example, kNN with EUC achieved an accuracy of 59% when trained with five dimensional HOOF data. This accuracy improved to 86% when five dimensional RMV data was used to train the kNN.

On the Weizmann dataset, when RMV is used in place of HOOF, the performance of the classifiers in some cases shows significant improvement. For example, while using kNN with EUC distance, the accuracy increases from 29% to 58% (when RMV is used instead of HOOF). Another example is when kNN is used with DTW+EUC, the accuracy increases from 49% to 58% when 12 dimensional RMV is used instead of 12 dimensional HOOF. For other cases, the performance of the classifiers are comparable. For example, the performance of kNN + NCM remains the same, at 57%, irrespective of the feature used.

## VII. FUTURE WORK

Although the performance of kNN on using RMV, either increased or was comparable, improvement in the performance

of HMM however was not very high. On the KTH dataset, with HOOF as the feature set, HMM achieved an recognition rate of 59% and with RMV feature set it obtained a recognition rate of 61%. Therefore, there is a scope of improving the performance of HMM further. Figures 9 and 10 shows the confusion matrices created by HMM when HOOF and RMV features are used respectively on the KTH dataset.

Predicted class								
		Вох	Clap	Wave	Jog	Run	Walk	
Actual class	Box	78	33	9	0	0	0	
	Clap	25	71	24	0	0	0	
	Wave	9	17	94	0	0	0	
	Jog	1	1	10	46	26	36	
	Run	0	0	9	33	58	20	
	Walk	0	2	6	23	11	78	

Fig. 9. Confusion matrix showing the performance of HMM when HOOF feature is used for activity recognition.

			Pred	icted class	5		
		Вох	Clap	Wave	Jog	Run	Walk
	Box	92	18	10	0	0	0
Actual class	Clap	20	92	8	0	0	0
	Wave	6	13	101	0	0	0
	Jog	1	3	0	21	50	45
	Run	0	3	0	19	79	19
	Walk	0	4	0	23	36	57

Fig. 10. Confusion matrix showing the performance of HMM when RMV feature is used for activity recognition.

The matrices reveal the following flaws while performing activity recognition:

- The classifier discriminates very well between static (boxing, clapping, waving) and mobile (jogging, walking and running) activities, less well between different static activities, and quite poorly between different mobile activities. Static activities in a video refer to activities where a subject is standing at one constant position. Mobile activities refer to activities where a subject is moving along the field of view throughout the video. Thus, while boxing, clapping and waving in the KTH dataset are static activities.
- The pace of mobile activities increases naturally from walking to jogging to running. Intuitively, one would expect that walking is misclassified as jogging more frequently than it is misclassified as running, and,

Future attempts to improve the performance of HMM can concentrate on solving these flaws.

Some other possible future works are as follows:

- RMV ignores direction of the flow vectors. For activity recognition direction is one of the important factors. Future attempts will try to incorporate direction into RMV.
- In RMV spatial information was incorporated by dividing a frame into sub-regions which are spatially correlated to each other. This division was done using a grid of resolution  $r \times s$ . The values of r and s were chosen as 10 and 20 respectively for the KTH dataset and as 18 and 16 respectively for the Weizmann dataset. However, these are not constant values and can be varied, i.e. a grid of different resolution can be chosen for the same dataset. Also, the values of r and s varies from one dataset to another. Future research can concentrate on studying the effect of varying r and s on a dataset and also on coming up with a more principled approach of selecting the values of r and s.
- Next, converting any feature set to their BOW representation and then using them with SVM for activity recognition is the state of the art in the activity recognition field. Thus, in future, RMV can also be converted it to its BOW representation and used with SVM to further assess its potentiality in video activity recognition. Further, the performance of RMV in its BOW form can be compared with the performance of HOOF in its BOW format. As RMV captures spatial information, spatial information is also added to its BOW representation. Similarly, as HOOF does not capture spatial information. A comparison between the BOW of RMV and BOW of HOOF will further illustrate the importance of spatial information.
- Finally, BOW itself lacks any spatial information and previously attempts have been made to incorporate spatial information in BOW. Some of these works include spatial pyramid [16] and CBOW [15]. A comparison of RMV in its BOW form with these works ([16] and [15]) on activity data may be another interesting research direction.

#### ACKNOWLEDGMENT

Shreeya Sengupta is supported by Vice Chancellor's Research Scholarship of University Of Ulster, UK.

#### REFERENCES

 J. R. R. Uijlings, A. Smeulders, and R. J. H. Scha, "What is the spatial extent of an object?" in *Computer Vision and Pattern Recognition*, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 770–777. [Online]. Available: http://www.huppelen.nl/publications/spatialExtentCvpr.pdf

- [2] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, 1999, pp. 1150–1157 vol.2. [Online]. Available: http://dl.acm.org/citation.cfm?id=850924.851523
- [3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *In CVPR*, 2005, pp. 886–893. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.177
- [4] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Computer Vision and Pattern Recognition*, 2009. CVPR 2009. *IEEE Conference on*, 2009, pp. 1932–1939. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2009.5206821
- [5] K. Brkić, A. Pinz, S. Šegvić, and Z. Kalafatić, "Histogram-based description of local space-time appearance," in *Proceedings of the 17th Scandinavian Conference on Image Analysis*, ser. SCIA'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 206–217.
- [6] C. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artificial Intelligence*, vol. 2012, p. 19 pages, 2012. [Online]. Available: http://dx.doi.org/10.5402/2012/376804
- [7] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 925–931.
- [8] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, Sep. 2008. [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0122-4
- [9] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*, sep 2008, pp. 995–1004.
- [10] F. M. Carrillo, A. Manzanera, and E. R. Castro, "A motion descriptor based on statistics of optical flow orientations for action classification in video-surveillance," in *Int. Conf. on Multimedia and Signal Processing* (CMSP'12), Shanghai, China, december 2012.
- [11] J. Pers, V. Sulic, M. Kristan, M. Perse, K. Polanec, and S. Kovacic, "Histograms of optical flow for efficient representation of body motion." *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2010.03.024
- [12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," 2013 IEEE Conference on Computer Vision and Pattern Recognition, vol. 0, pp. 1–8, 2008.
- H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the 2011 IEEE Conference* on Computer Vision and Pattern Recognition, ser. CVPR '11.
   Washington, DC, USA: IEEE Computer Society, 2011, pp. 3169–3176.
   [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995407
- [14] M. Grundmann, F. Meier, and I. A. Essa, "3d shape context and distance transform for action recognition." in *ICPR*. IEEE, 2008, pp. 1–4.
- [15] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *Circuits and Systems for Video Technology*, *IEEE Transactions on*, vol. 21, no. 4, pp. 381–392, 2011. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2010.2041828
- [16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition*, 2006 IEEE Computer Society Conference on, vol. 2, 2006, pp. 2169–2178. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.68
- [17] K. Brkic, S. Rasic, A. Pinz, S. Segvic, and Z. Kalafatic, "Combining spatio-temporal appearance descriptors and optical flow for human action recognition in video data," *CoRR*, vol. abs/1310.0308, 2013.
- [18] 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, 2004, pp. 32–36 Vol.3. [Online]. Available: http: //dx.doi.org/10.1109/ICPR.2004.747
- [19] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.70711

- [20] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision (ijcai)," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81)*, April 1981, pp. 674–679.
- H. Wang, "Nearest neighbors by neighborhood counting," *IEEE Trans.* Pattern Anal. Mach. Intell., vol. 28, pp. 942–953, June 2006. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2006.126
- [22] P. Senin, "Dynamic Time Warping Algorithm Review," Department of Information and Computer Sciences, University of Hawaii, Honolulu, Hawaii 96822, Tech. Rep. CSDL-08-04, Dec. 2008.
- [23] D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences," *Commun. ACM*, vol. 18, no. 6, pp. 341–343, Jun. 1975. [Online]. Available: http://doi.acm.org/10.1145/360825. 360861
- [24] H. Wang, "All common subsequences," in *Proceedings of the 20th international joint conference on Artifical intelligence*, ser. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 635–640.
- [25] K. Murphy, "Hidden Markov Model (HMM) toolbox for matlab," 1998.