Saliency-Informed Spatio-Temporal Vector of Locally Aggregated Descriptors and Fisher Vector for Visual Action Recognition

Zheming Zuo
zheming.zuo@northumbria.ac.uk
Daniel Organisciak
daniel.organisciak@northumbria.ac.uk
Hubert P. H. Shum
hubert.shum@northumbria.ac.uk
Longzhi Yang
longzhi.yang@northumbria.ac.uk

Department of Computer and
Information Sciences
Northumbria University
Newcastle upon Tyne, NE1 8ST, UK

Abstract

Feature encoding has been extensively studied for the task of visual action recognition (VAR). The recently proposed super vector-based encoding methods, such as the Vector of Locally Aggregated Descriptors (VLAD) and the Fisher Vector (FV), have significantly improved the recognition performance. Despite of the success, they still struggle with the superfluous information that presents during the training stage, which makes the methods computationally expensive when applied to a large number of extracted features. In order to address such challenge, this paper proposes a Saliency-Informed Spatio-Temporal VLAD (SST-VLAD) approach which selects the extracted features corresponding to small amount of videos in the data set by considering both the spatial and temporal video-wise saliency scores; and the same extension principle has also been applied to the FV approach. The experimental results indicate that the proposed feature encoding scheme consistently outperforms the existing ones with significantly lower computational cost.

1 Introduction

Visual Action Recognition (VAR) is a growing research field applicable to application areas including human-computer interaction [22], video [26], motion analysis [17], clinical science [29], pervasive health-care (*e.g.* fall detection) [4], sports analysis, and gaming amongst others. Vector of Local Aggregated Descriptors (VLAD) and Fisher Vector (FV) are common feature encoding methods and their efficacy has been proven by a multitude of tasks within the computer vision domain[2, 4, 9, 13, 20, 24, 25, 27, 28, 29]. They attain high performance for both first-person[29] and third-person action recognition tasks[4, 9, 13, 20, 25, 27]. VLAD is usually less computationally demanding than FV, but this comes with the cost of often being less accurate in terms of classification precision. One downside of VLAD and

FV is they do not consider any temporal information. Additionally, neither algorithm is able to cope with the presence of redundancy that is embedded within the training datasets.

Action recognition suffers from many of the common challenges often seen in the field of computer vision including background clutter, diverse lighting conditions, camera motion, and occlusion [4, 29]. In particular, the camera motion problem is particularly associated with egocentric action recognition [29]. Even relatively small movements can result in a significant change in the background. In extreme cases, this can result in blurred video frames or missing a part of the action which severely impacts the ability of the existing classification approaches.

Superfluous information embedded within the action recognition data set can degrade the quality of the generated code-book during feature encoding. For instance, if action recognition techniques are used to detect falls and one of the training videos is thirty seconds long while the fall lasts only two seconds, then the majority of this video is not relevant to the task at hand. This could negatively impact the final result. In order to lessen the computational cost, extracted features are commonly subject to a random sub-sampling strategy [7] before VLAD or FV feature encoding is applied. However, this strategy is sub-optimal due to the uncertainty that it introduces.

In this work, a solution is identified which handles both of the aforementioned problems. In particular, a family of feature encoding schemes are proposed on the basis of the calculation of video-wise spatio-temporal saliency scores. The redundancy is eliminated using such schemes by only considering the video clips with the highest saliency and then extract features for VLAD and FV feature encoding only from these ones. Because the features are extracted only from the more apposite videos, there are fewer initial extracted features. This leads to very efficient feature encoding with no need to introduce any random sub-sampling.

The framework of the propsoed appraoch is outlined in Figure 1. In this approach, the spatial saliency and temporal saliency for each video are calculated first, which are then combined to obtain a spatio-temporal score. From this, only the most salient videos are selected for code-book generation so that the learned dictionary is more pertinent to the key elements of the data set and the superfluous data are generally ignored.

The main contributions of this work are: 1) proposing two families of spatio-temporal feature encoding schemes for both first- and third-person action recognition; 2) constructing spatio-temporal video-wise saliency scores in order to help eliminate redundancy in the data set and to speed up the code-book generation process; 3) conducting extensive experiments on three different data sets with five different feature extraction methods. In no case did the standard feature encoding methods outperform the proposed saliency-based ones according to the experimentation, which demonstrates the power of the proposed approach.

2 Related Work

Super-vector based feature encodings such as VLAD [12] and FV [21] have been widely employed in the context of VAR tasks [7, 20, 25, 27, 29]. VLAD essentially aggregates the residuals between feature descriptors and visual words. However, VLAD is lacking the ability to capture the temporal information during the encoding phase, Duta *et al.* [8] proposed the Spatio-Temporal VLAD (ST-VLAD) for addressing this by incorporating the VLAD with 3-D feature positions.

There are numerous ways to extract the features for VLAD and FV encoding schemes. Dalal and Triggs [6] proposed Histograms of Oriented Gradients (HOG) to capture the spatial

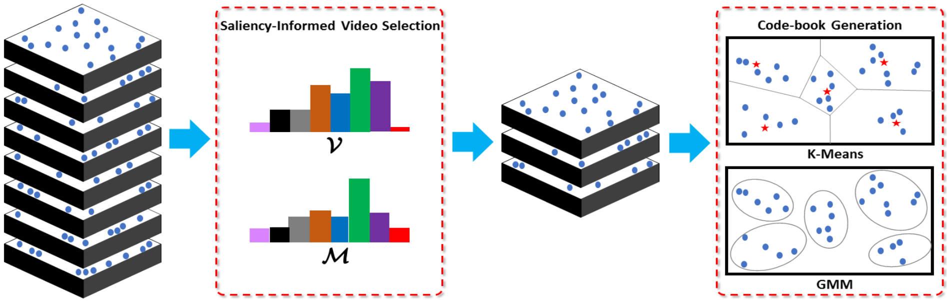

Figure 1: A visualisation of the proposition. The most relevant videos in the data set are selected which results in fewer extracted features thus feature encoding can be performed with less computational cost.

or static appearance information of human actions, Duta *et al.* [7] introduced Histograms of Motion Gradients (HMG) which conducts simple temporal derivative prior to the HOG process, Laptev *et al.* combined HOG with optical flow (HOF) [16] to aggregate 2D optical flow responses, Dalal *et al.* [6] introduced Motion Boundary Histograms (MBH) which remove constant motion to handle camera motion by calculating horizontal and vertical derivatives on the calculated optical flow from HOF separately (*i.e.* MBHx and MBHy). Zuo *et al.* [29] developed gaze-informed counterparts to the aforementioned features to be employed for recognising first-person actions.

Compared with FV, VLAD ignores the complementary high-order statistics of the extracted feature descriptors, [18] proposed high-order VLAD that works jointly with the supervised dictionary learning. FV usually achieves better precision than VLAD, [19] constructed stacked FV with more semantic information extracted using a hierarchical structure. However, practically, in the case when coping with large-scale visual action recognition data sets, it is usually not affordable to use all the extracted features for code-book generation. Thus, randomly selecting a small portion of feature descriptors is a possible solution [7]. However, random down-sampling cannot guarantee to provide a representative sample because uncertainty is introduced during this random selection. In order to avoid this, all generated features must be used which may result in longer experimental (including feature encoding) time despite the more competitive performance.

Though intensive efforts have been made to enhance the performance of feature representations, one significant challenge which is still a long way from being solved is the large computational cost. At present, the aforementioned random down-sampling strategy is commonly used. In this paper, two families of extensions of VLAD and FV are proposed to reduce the computational cost in a more deterministic fashion while guaranteeing the classification performance. More specifically, the concept of saliency is used to guide the process of selecting the most important and discriminative videos within the data set for code-book generation in the encoding phase.

3 Saliency-Informed Spatio-Temporal Feature Encoding

The framework of the proposed saliency-informed spatio-temporal feature encoding schemes is visualised in Figure 2.

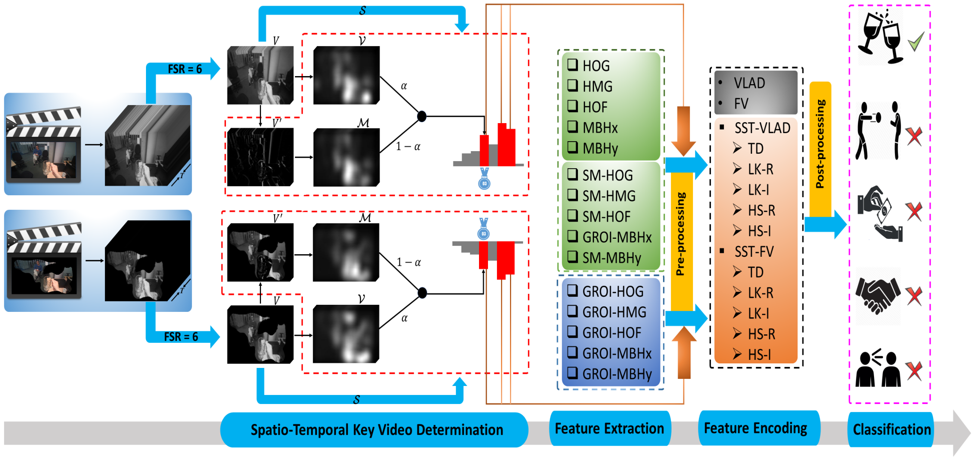

Figure 2: The framework of Saliency-informed Spatio-Temporal (SST) feature encodings for visual action recognition.

3.1 VLAD and FV: A Revisit

Generally, VLAD starts with the calculation of centroids in the feature space using k-means clustering algorithm, which is followed by the aggregation of features using the calculated centroids. This results in encoded features of dimension $k \times d$, where k is the number of centroids (*i.e.* visual words) and d is the dimensionality of the encoded feature vector. Therefore, VLAD can be viewed as a simplified non-probabilistic version of FV, and thus VLAD is practically faster but with relatively poor performance.

The code-book generation method used in VLAD [12] is k-means. Given a set of pre-processed local features $\mathcal{F} = \{f_1, \dots, f_n\}$ where n is the number of local features (*i.e.* number of videos in visual action recognition), f_i with $1 \le i \le n$ denotes the i-th local feature. Let $\mathcal{C}$ be the set of k clusters, $\mathcal{C} = \{c_1, \dots, c_k\}$, c_j with $1 \le j \le k$ is a prototype associated with the j-th cluster. The k-means algorithm generates the clusters via the following objective function:

$$\min_{\{\psi_{ij}, c_j\}} \sum_{i=1}^{n} \sum_{j=1}^{k} \psi_{ij} \|f_i - c_j\|_2^2, \tag{1}$$

where ψ_{ij} is a Boolean indicator variable setting to either 1 (when local descriptor f_i is assigned to cluster j) or 0 (otherwise). In VLAD, the k-means algorithm transforms each local feature descriptor from the feature space to code-word by performing such a hard assignment.

For FV, Gaussian Mixture Models (GMM) are used to convert the local feature set to the code-word by performing soft assignment for each local feature. Suppose L clusters are required, denote the parameters of the GMM, u_λ, by $\lambda = \{w_l, \mu_l, \Sigma_l; l = 1, \dots, L\}$ where w_l is the mixture weight, μ_l is the mean vector and Σ_l is the covariance matrix. Then

$$u_\lambda(x) = \sum_{l=1}^{L} w_l u_l(x), \tag{2}$$

where u_l is the Gaussian l.

These approaches are sub-optimal for action recognition as they do not incorporate any temporal information. Additionally, a certain degree of redundancy is inherent in most real-world visual action recognition data sets. Encoding features based on this redundancy is computationally wasteful and reduces performance. Moreover, when computing these super-vectors, the most frequent method employed to extract features to be encoded relies on a random down-sampling strategy. This method may not guarantee that the randomly selected features are capable of providing discriminative and continuous motion clues for later stages of the pipeline, *i.e.* feature encoding and classification.

3.2 Saliency-Informed ST-VLAD and ST-FV

In this section, the proposed approach is introduced which aims to address the aforementioned redundancy and uncertainty by specifically determining the key videos from which to extract features for code-book generation.

3.2.1 Saliency-Informed Key Video Determination

The proposed approach utilises spatio-temporal video-wise saliency maps to inform and guide the process of code-book generation during feature encoding. To generate the saliency maps, the image signature [11] is used due to its computational efficiency. Motion information is also captured via optical flow (*e.g*. Lukas-Kanade [3], Horn-Schunck [10]) or a simple temporal derivative. The optical flow calculation returns complex-valued vectors whereby the real and imaginary components can be considered separately in determining the temporal saliency score.

For a video, V, consisting of N frames $\{F_1,\ldots,F_N\}$ the temporal derivative is computed as

$$F' = F_i - F_{i-1}, \tag{3}$$

where $i = 2,\ldots,N$.

To obtain the saliency maps, the videos are firstly converted from the RGB space to LAB space. Given a video in LAB space, V, consisting of N frames, $\{F_1,\ldots,F_N\}$, across the L, A, and B colour channels, the video signature $\mathcal{V}$ is calculated:

$$\mathcal{V} = \sum_{n=1}^{N}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\kappa_g * \left[\frac{1}{3}\sum_{c\in\{L,A,B\}}\mathrm{IDCT}(\mathrm{sign}(\mathrm{DCT}(F_i^c)))\right]^{\circ 2}\right), \tag{4}$$

where $[\cdot]^{\circ 2}$ denotes the Hadamard power of order two, κ_g is the Gaussian kernel, $\mathrm{DCT}(\cdot)$ is the Discrete Cosine Transform, $\mathrm{IDCT}(\cdot)$ is its inverse form, and $\mathrm{sign}(\cdot)$ maps positive numbers to $+1$ and negative numbers to -1.

The motion information is also incorporated in a similar way. the motion information V' of a video V is obtained by applying either Eq. (3) or optical flow calculation methods such as Lucas-Kanade or Horn-Schunck. V' consists of $N-1$ motion images, $\{F'_1,\ldots,F'_{N-1}\}$, across the L, A, and B colour channels. The motion signature $\mathcal{M}$ is proposed:

$$\mathcal{M} = \sum_{n=1}^{N-1}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\kappa_g * \left[\frac{1}{3}\sum_{c\in\{L,A,B\}}\mathrm{IDCT}(\mathrm{sign}(\mathrm{DCT}(F'^{c}_{i})))\right]^{\circ 2}\right). \tag{5}$$

Finally, a weight, α is introduced, to give the spatial saliency (*i.e.* video signature) more importance than the temporal saliency (*i.e.* motion signature) in calculating the overall saliency score, $\mathcal{S}$:

$$\mathcal{S} = \alpha\mathcal{V} + (1-\alpha)\mathcal{M}. \tag{6}$$

In the experiments, it is empirically determined that $\alpha = 0.75$ provides optimal performance. Once a saliency score for each video has been calculated, only the videos within the top 30% are considered for feature encoding.

4 Experiments

4.1 Data Sets and Settings

Three visual action recognition data sets were chosen for evaluating the proposed SST-VLAD and SST-FV families of feature encoding schemes. (a) **UNN-GazeEAR data set** [29] is an egocentric and interactive action recognition data set that consists of five action categories, each of which includes ten video clips and each of which is with different time durations ranging from 2 to 11 seconds. (b) **UNN-6 data sets**[1] [4] which comprises of two sub-sets: colour (namely **UNN6_Color**) and infra-red (*i.e.* **UNN6_IR**). These data sets are third-person fall and similar daily activity action data sets that contain six classes of human actions with six videos per class. For comparison purposes, the two categories of local features are adopted, including (HOG, HMG, HOF, MBHx, and MBHy), and their gaze-region-of-interest based counterparts. All the experiments were carried out using a HP workstation with Intel® Xeon™ E5-1630 v4 CPU @ 3.70 GHz with 64 GB RAMs.

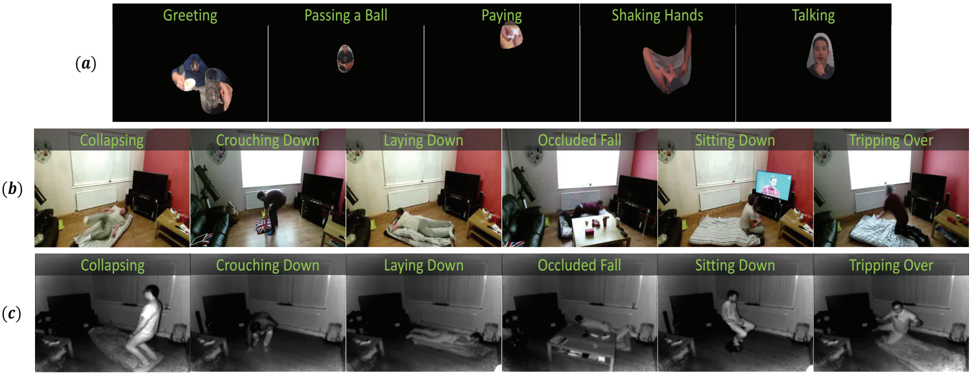

Figure 3: Sample frames from video sequences of (a) UNN-GazeEAR with GROI preprocessed, (b) UNN-6 Colour, and (c) UNN-6 Infrared data sets.

In the feature extraction phase, the block size is set to 4-by-4 spatial pixels by 6 frames and the frame sampling rate (FSR) is valued as 6 frames. Then, in the pre-processing stage, the extracted local features are pre-processed by RootSIFT [1] normalisation technique which is followed by reducing the features to 72 dimensions using PCA. In the post-processing stage, the power normalisation plus $\ell 2$ normalisation (PN$\ell 2$) is used to further normalisation the encoded features, in which the power parameter is set to 0.5. The number

[1] http://computing.northumbria.ac.uk/staff/FGPD3/unn6-2017/

UNN-GazeEAR		GROI-HOG			GROI-HMG			GROI-HOF			GROI-MBHx			GROI-MBHy		
		8	16	32	8	16	32	8	16	32	8	16	32	8	16	32
VLAD		90.36	91.58	94.10	91.16	90.12	91.72	91.24	89.88	92.08	90.40	91.84	91.46	89.50	92.16	93.48
SST-VLAD	TD	90.40	92.80	94.92	92.14	91.12	**93.20**	91.32	90.48	92.20	90.98	92.62	92.28	90.84	92.98	94.30
	LK-R	92.14	92.38	94.12	91.90	92.02	92.74	92.08	90.20	92.24	91.26	91.98	91.88	90.60	94.22	94.02
	LK-I	92.04	**93.04**	**95.30**	91.78	92.44	92.84	91.38	90.30	**93.30**	90.46	92.26	**93.40**	**91.74**	93.44	**94.80**
	HS-R	92.10	92.30	94.68	91.86	91.86	92.46	**93.44**	90.06	92.38	90.96	**93.92**	92.06	89.72	**94.72**	94.06
	HS-I	**92.72**	92.56	94.48	**92.32**	**92.48**	93.00	91.58	**92.70**	93.00	**92.28**	93.12	92.72	90.40	92.48	94.62
FV		93.30	94.44	96.46	91.00	92.78	94.96	90.20	93.06	91.94	88.10	88.74	91.18	88.24	90.46	91.86
SST-FV	TD	**94.52**	94.94	**96.88**	92.36	94.78	**96.40**	90.74	93.08	92.78	90.24	89.64	92.42	**92.44**	92.70	93.06
	LK-R	94.04	94.74	96.68	91.06	**95.10**	95.28	90.78	**94.98**	93.26	89.70	**92.18**	92.06	90.56	91.40	92.74
	LK-I	93.50	95.24	96.76	91.20	93.68	95.90	92.00	93.54	92.70	89.88	90.94	92.12	90.38	**92.72**	**93.66**
	HS-R	93.88	95.12	96.50	**92.70**	94.12	95.84	91.84	93.20	92.72	90.24	89.00	92.12	88.98	91.84	93.14
	HS-I	94.48	**95.44**	96.66	92.44	94.84	95.80	**92.62**	94.16	**93.64**	**90.78**	91.94	**93.76**	90.28	90.94	93.38

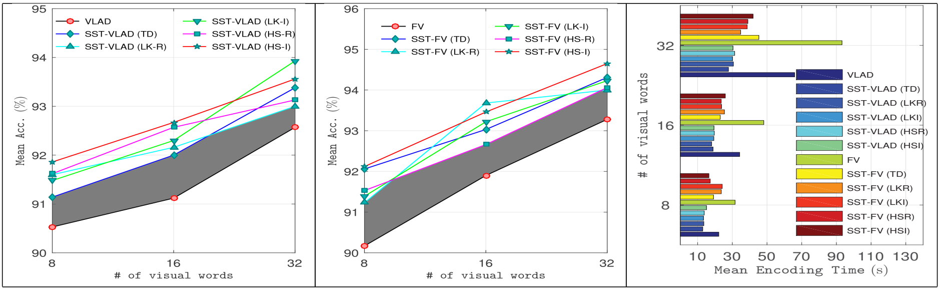

Table 1: UNN-GazeEAR - Accuracy generated by varying the number of visual words in different feature encoding schemes. The best performance is marked in green while the worst performance is marked in red. The grey region demonstrates the difference between the standard FV and VLAD encoding and the worst results obtained by the proposed method.

of nodes in the hidden layer of the neural network is 20. Due to more competitive performance, the Horn-Schunck method is particularly adopted in HOF, MBHx, and MBHy for optical flow calculation in the feature extraction stage [25].

For VLAD, the feature encoding time reported here includes the time to perform the k-means algorithm and the hard assignment. For Fisher vectors, it encompasses the GMM clustering and soft assignment. The performance of the proposed encoding schemes was evaluated with different dictionary sizes of 8, 16, and 32 to ensure that the findings are consistent. The relatively small sizes of the data sets used mean they are not suitable for evaluation on longer dictionaries.

The accuracy reported throughout this section is generated by performing each experiment one hundred times and then obtaining the mean of these experiments. Overall, with five different feature extraction approaches, five methods to calculate temporal saliency and three different dictionary sizes, 150 results are presented for each data set to thoroughly test the proposed method and demonstrate its efficacy.

4.2 The Effect of Various Motion Calculation based Key Videos Determination

The quantitative results are presented in Tables 1, 2 and 3. For intuitive comparison between the different methods of calculating temporal saliency, the lowest score (red) and highest score (green) are highlighted for each dictionary size and feature extraction method. It can

UNN6: Color Set		HOG			HMG			HOF			MBHx			MBHy		
		8	16	32	8	16	32	8	16	32	8	16	32	8	16	32
VLAD		88.39	90.03	90.78	88.03	89.00	87.83	82.50	86.11	88.56	89.42	90.53	89.11	88.39	85.75	87.81
SST-VLAD	TD	**93.97**	**93.28**	92.39	88.89	**89.92**	**89.89**	**86.39**	**88.53**	**89.78**	**91.31**	90.69	**90.58**	**89.28**	86.81	88.33
	LK-R	93.58	93.00	91.92	89.06	89.53	88.94	83.64	87.56	88.89	90.22	**91.42**	90.03	88.75	85.97	**88.75**
	LK-I	93.83	92.08	**93.11**	88.58	89.08	89.36	83.97	87.42	88.94	90.61	90.75	90.47	88.64	87.14	88.42
	HS-R	93.31	92.89	**93.11**	**89.86**	89.22	89.42	83.94	87.94	89.28	90.44	90.56	90.25	88.56	87.06	88.53
	HS-I	93.36	91.00	93.00	88.47	89.61	89.33	84.36	87.78	89.33	90.19	90.78	90.03	88.75	**88.06**	88.69
FV		91.39	88.92	90.78	91.56	91.81	92.19	83.56	90.19	89.22	92.72	91.67	91.67	92.64	92.11	90.78
SST-FV	TD	92.00	91.75	92.83	93.11	92.42	92.36	86.83	91.33	90.47	92.81	92.56	**92.64**	92.94	**92.92**	**92.75**
	LK-R	**93.31**	**93.92**	92.11	**93.36**	**93.50**	92.97	87.14	**91.67**	90.47	92.92	92.64	92.61	92.97	92.64	92.31
	LK-I	93.14	92.89	**93.61**	92.39	92.50	92.83	85.94	91.31	89.92	**93.42**	92.39	91.81	92.81	92.19	92.08
	HS-R	93.19	93.53	93.14	92.67	92.53	92.53	87.64	90.56	**91.92**	92.78	**92.75**	92.50	**93.11**	92.75	92.44
	HS-I	92.78	93.53	93.19	92.25	92.61	**93.14**	**89.58**	91.19	91.64	92.81	91.92	91.69	92.97	92.83	91.94

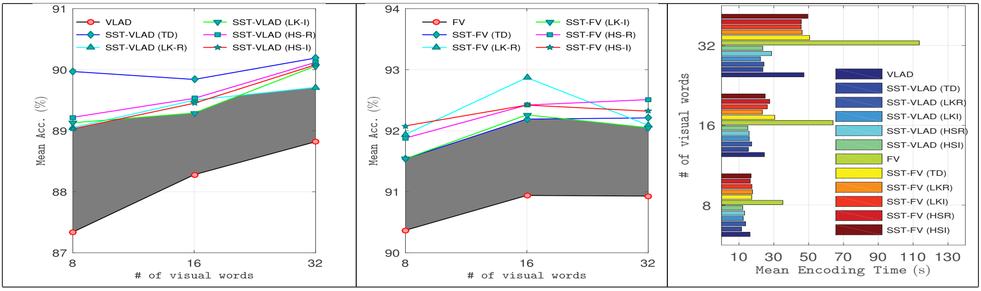

Table 2: UNN6_Color Set - Accuracy generated by varying the number of visual words in different feature encoding schemes. The best performance is marked in green while the worst performance is marked in red. The grey region demonstrates the difference between the standard FV and VLAD encoding and the worst results obtained by the proposed method.

be seen clearly that in all cases, the ordinary VLAD and FV have the weakest performance whilst no temporal saliency calculation method stands out as definitively outperforming the others.

The left and centre figures below each table show the mean accuracy of each of the six encoding methods across all different feature extraction approaches. The rightmost figure below each table shows the mean encoding time for each method across all feature extraction techniques. The eliminated redundancy in the data set significantly improves the efficiency of the code-book generation. In most cases, the process is approximately twice as fast as the standard VLAD and FV.

The averaged results across all different feature extraction techniques and dictionary lengths are presented in Table 4 to facilitate the comparison. No temporal saliency calculation technique stands out by a significant margin but the imaginary component of Horn-Schunck performs strongest in three of the six tasks and has competitive performance in the remaining three.

5 Conclusion

Two potential challenges in visual action recognition have been identified in this paper which arise when using a super-vector based feature encoding approach: 1) redundancy embedded in the data set can weaken categorisation precision; 2) the large number of extracted features result in inefficient feature embedding from the perspectives of space and time complexi-

UNN6: IR Set		HOG			HMG			HOF			MBHx			MBHy		
		8	16	32	8	16	32	8	16	32	8	16	32	8	16	32
VLAD		97.97	98.69	99.53	92.11	94.36	94.86	94.39	94.14	95.67	95.69	96.17	95.22	93.44	93.81	94.69
	TD	**99.31**	99.00	99.83	**94.78**	94.86	96.19	96.56	95.42	96.56	96.97	97.39	95.72	94.89	94.22	**96.19**
	LK-R	98.78	99.50	**99.94**	93.00	95.00	97.06	94.58	95.03	96.00	96.19	97.17	96.06	**95.31**	94.64	95.47
SST-VLAD	LK-I	98.81	**99.75**	99.83	92.44	**97.53**	96.44	**96.58**	95.47	95.86	97.19	97.42	95.47	93.58	**94.83**	95.81
	HS-R	98.28	99.69	99.64	94.08	95.72	95.81	94.61	94.78	95.92	95.75	97.11	**97.78**	94.72	94.31	96.03
	HS-I	98.92	99.64	99.83	93.08	96.17	**97.08**	95.50	**95.67**	**96.78**	**97.28**	**97.81**	97.08	94.31	**94.83**	95.83
FV		99.03	99.69	99.78	98.28	98.08	95.03	95.78	96.00	96.14	97.89	96.33	96.42	94.75	95.00	96.28
	TD	99.69	99.94	99.89	98.56	**99.03**	96.28	96.75	96.53	**97.31**	99.00	97.94	**98.47**	96.14	95.53	97.00
	LK-R	99.86	99.72	**100.00**	**98.81**	98.39	96.53	97.17	97.03	97.00	97.92	96.47	97.53	95.19	**98.03**	97.06
SST-FV	LK-I	**99.92**	99.78	**100.00**	98.47	98.28	**97.89**	96.61	**97.31**	96.44	98.08	**98.47**	97.11	96.22	96.11	97.50
	HS-R	99.75	**99.97**	99.86	98.64	98.86	97.14	**98.42**	96.92	97.17	98.61	97.56	97.03	**96.61**	96.92	97.47
	HS-I	99.78	99.92	99.86	98.64	98.75	96.22	96.47	97.00	96.31	**99.19**	97.53	96.81	95.53	96.03	**97.72**

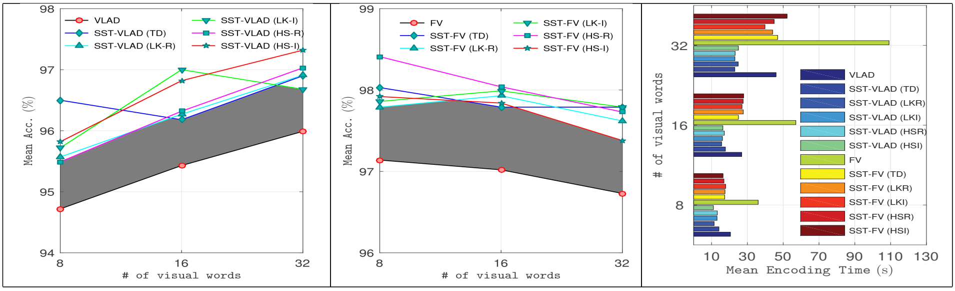

Table 3: UNN6_IR Set - Accuracy generated by varying the number of visual words in different feature encoding schemes. The best performance is marked in green while the worst performance is marked in red. The grey region demonstrates the difference between the standard FV and VLAD encoding and the worst results obtained by the proposed method.

ties. An salicency-based approach has therefore been proposed in this paper such that only the most relevant videos via video-wise saliency are sampled in order to solve both of these problems. The experiments demonstrated that video frames can be discriminated based upon their spatio-temporal saliency score for first- and third-person action recognition scenarios, which can significantly boost the performance of VLAD and FV whilst simultaneously making the process considerably more efficient. In particular, it has been proven that focusing primarily on videos which are sparsely packed with relevant information consistently leads to better categorisation accuracy.

There is a comprehensive amount of future work in the form of extensions of these initial proposal and experiments. The work can be further evaluated by applying the proposed

Overall Accuracy		TD	LK-R	LK-I	HS-R	HS-I
	UNN-GazeEAR	92.17	92.25	92.57	92.44	**92.70**
SST-VLAD	UNN6_Color	**90.00**	89.42	89.49	89.62	89.52
	UNN6_IR	96.53	96.25	96.47	96.28	**96.65**
	UNN-GazeEAR	93.13	92.97	92.95	92.75	**93.41**
SST-FV	UNN6_Color	91.98	**92.30**	91.95	92.27	92.27
	UNN6_IR	97.87	97.78	97.88	**98.06**	97.72

Table 4: Mean results of different methods of calculating temporal saliency across all different feature extraction techniques and code-book sizes on all the data sets employed in this work. Bold results indicate highest performance on a particular data set.

framework to more challenging, larger scale data sets such as UCF50 [23], UCF101 [14] and HMDB51 [15]. Intuitively, decreasing the frame sampling rate (leads to more dense video representations) will improve performance, though it will increase the time taken for code-book generation. It is therefore interesting to explore using a spatio-temporal saliency sliding window to extract the key segments of a video as a method to dispose of superfluous data.

References

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[2] R. Arandjelović and A. Zisserman. All about vlad. In *CVPR*, 2013.

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[4] R. Cameron, Z. Zuo, G. Sexton, and L. Yang. A fall detection/recognition system and an empirical study of gradient-based feature extraction approaches. In *UKCI*, 2017.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[7] I. C. Duta, J. R. R. Uijlings, T. A. Nguyen, K. Aizawa, A. G. Hauptmann, B. Ionescu, and N. Sebe. Histograms of motion gradients for real-time video classification. In *CBMI*, 2016.

[8] I. C. Duta, B. Ionescu, K. Aizawa, and N. Sebe. Spatio-temporal vlad encoding for human action recognition in videos. In *MMM*, 2017.

[9] I. C. Duta, J. R. R. Uijlings, B. Ionescu, K. Aizawa, A. G. Hauptmann, and N. Sebe. Efficient human action recognition using histograms of motion gradients and vlad with descriptor shape information. *Multimedia Tools and Applications*, pages 1–28, 2017.

[10] B. K. Horn and B. G Schunck. Determining optical flow. *Artificial Intelligence*, 17 (1-3):185–203, 1981.

[11] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):194–201, 2012.

[12] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1704–1716, 2012.

[13] V. Kantorov and I. Laptev. Efficient feature extraction, encoding and classification for action recognition. In *CVPR*, 2014.

[14] S. Khurram, R. Z. Amir, and S. Mubarak. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR, abs/1212.0402*, 2012.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011.

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[17] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *ICCV*, 2009.

[18] X. Peng, L. Wang, Y. Qiao, and Q. Peng. Boosting vlad with supervised dictionary learning and high-order statistics. In *ECCV*, 2014.

[19] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, 2014.

[20] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109 – 125, 2016.

[21] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[22] S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015.

[23] K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[24] K. Simonyan, O. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In *BMVC*, 2013.

[25] J. R. R. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe. Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. *International Journal of Multimedia Information Retrieval*, 4(1): 33–44, 2015.

[26] S. Vishwakarma and A. Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10):983–1009, 2013.

[27] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[28] S. Yan, X. Xu, D. Xu, S. Lin, and X. Li. Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification. In *ECCV*, 2012.

[29] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu. Gaze-informed egocentric action recognition for memory aid systems. *IEEE Access*, 6:12894–12904, 2018.