Proceedings of IEEE Int'l Conference on Image Processing (ICIP-2016), Phoenix, AZ, USA, Sep. 2016.

INTRODUCING TEMPORAL ORDER OF DOMINANT VISUAL WORD SUB-SEQUENCES FOR HUMAN ACTION RECOGNITION

N. Kardaris, V. Pitsikalis, E. Mavroudi, P. Maragos

School of ECE, National Technical University of Athens, 15773 Athens, Greece
{vpitsik,maragos}@cs.ntua.gr, nick.kardaris@gmail.com, emavrou1@jhmi.edu

ABSTRACT

We present a novel video representation for human action recognition by considering temporal sequences of visual words. Based on state-of-the-art dense trajectories, we introduce temporal bundles of dominant, that is most frequent, visual words. These are employed to construct a complementary action representation of ordered dominant visual word sequences, that additionally incorporates fine grained temporal information. We exploit the introduced temporal information by applying local sub-sequence alignment that quantifies the similarity between sequences. This facilitates the fusion of our representation with the bag-of-visual-words (BoVW) representation. Our approach incorporates sequential temporal structure and results in a low-dimensional representation compared to the BoVW, while still yielding a descent result when combined with it. Experiments on the KTH, Hollywood2 and the challenging HMDB51 datasets show that the proposed framework is complementary to the BoVW representation, which discards temporal order.

Index Terms— visual human action recognition, bag-of-visual-words, video representation, temporal sequences, local sub-sequence alignment.

1. INTRODUCTION

Since the Bag-of-Visual-Words (BoVW) [1] was introduced, its combinations with variants of spatio-temporal feature descriptors [2] have become popular for visual human action recognition [3], and still draw attention [4, 5]. Despite its effectiveness, there are several issues that are not dealt with, opening the way for supplementary advancements: these issues are related to the spatio-temporal information [6] that is disregarded at the encoding stage, where features are quantized and statistics over their distribution are aggregated into vector representations. Motivated by the lack of temporal information, we add temporal structure to action sub-sequences, that enriches the actions' description in an intuitive way and is further shown to increase performance. After all, temporal structure is inherent to many cases of everyday human actions included in competitive action video datasets [7].

Actions which are roughly symmetric in time, such as stand-up versus sit-down, are expected to have similar statistics, but different dynamics. Thus, temporal context is essential for action recognition which consist of similar sub-actions, but in different order. Such an example is shown in Fig. 1, where actions "stand" (first row) and "sit" (second row) are depicted along with their corresponding BoVW histograms. The BoVW representation aggregates the occurrence frequency of each visual word in a single histogram, discarding temporal information. As a consequence, actions with similar visual words' frequency but different distribution of these words across time would be incorrectly classified to the same class. Instead the proposed approach retains the visual words' temporal order.

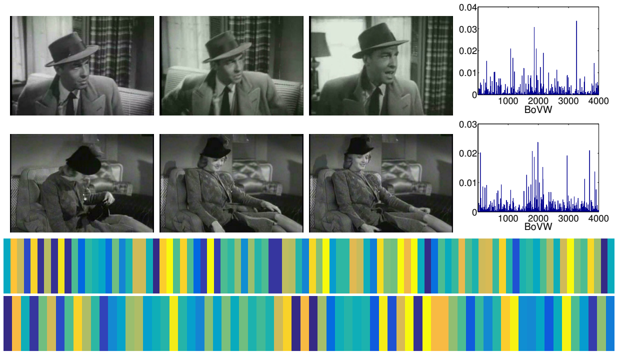

Fig. 1. Sample HMDB51 [7] sequences from actions "stand" (first row) and "sit" (second row) along with their respective bag-of-visual-words representation. BoVW ignores essential temporal information. Third and fourth row: sequential representation of the two actions by employing the proposed scheme, providing enriched temporal information.

In this work, we face the above by introducing a new approach that incorporates temporal sequential structure. We encode dense trajectories features [8] as a sequence of most frequently occurring visual words. Each video is short-time processed using a sliding window, yielding visual word bundles. Visual words within each bundle are temporally ordered and the most dominant ones, in terms of frequency, are retained (Sec. 3.1). The concatenation of these visual word subsequences results in a Sequence of Dominant Visual Words (SoDVW) representing the video. We introduce temporal information at two levels: First, these dominant visual word subsequences carry by construction fine-grained temporal information. Second, by concatenating them we incorporate the temporal subsequence order into the global action representation. Pairwise similarity of temporal sequences is measured by employing local alignment of action sub-sequences taking into consideration quantitative metrics of the feature space (Sec.3.2). The pairwise alignment similarity scores are incorporated at the the support vector machine kernel (Sec. 3.3). The overall framework is evaluated on the KTH [9], Hollywood2 [10] and HMDB51 [7] datasets, leading to consistent relative improvements of up to 5% (Sec. 4).

This research work was supported by the European Union under the project MOBOT with grant FP7-ICT-2011.2.1-600796 and in part under the EU project I-SUPPORT with grant H2020-643666.

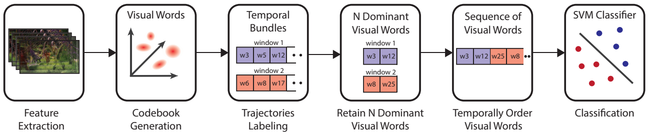

Fig. 2. Overview block diagram of the proposed approach.

2. RELATED WORK

Human action recognition is an active research area. A variety of demanding datasets [11, 7], feature descriptors [12, 13, 8] and encoding methods [14, 15, 16] have been introduced in an attempt to face several challenges [17]. Approaches span several directions, such as "deep" architectures [18, 19, 20] and the famous Bag-of-Visual-Words (BoVW) paradigm [21, 2, 3], as well as, one of the top performing approaches, the dense trajectories [4, 8]. Apart from the efficacy of the above, important aspects are ignored, such as spatio-temporal information at the level of the video representations [6, 22].

Other attempts deal with temporal information within BoVW. Spatio-temporal pyramids [12] integrate spatial-temporal information by decomposing the whole video into spatio-temporal subvolumes and computing a BoVW in each. Glaser *et al.* [23] consider time-enriched feature vectors introduced by [24] to form visual word sequences and build "video parts" by aggregating visual words (VW) at consecutive frames. Cheng *et al.* [25] model temporal relations between action parts using a subset of Allen's relations. Agustí *et al.* [5] implicitly integrate temporal relations within the BoVW model, by encoding VWs co-occurrences for several time displacements. Others [26] use correlograms to model the local spatio-temporal relations between pairs of VW by computing histograms of locally co-occuring words. Bettadapura *et al.* [27] augment BoVW with histograms of n-grams to describe relations between temporally overlapping events. More recently, Nagel et al. [28] model the temporal relationship among frames using a HMM and a Fisher vector variant. Fernando et al. [29] model the action evolution in time by learning the frames' temporal order using ranking machines. Finally, others incorporate hidden Markov models [30, 31, 32]. As opposed to our work, most of the above enrich the BoVW in several aspects, e.g. computing statistics among VWs or exploiting video parts.

The work of [33] bears similarity to ours. They use SIFT features and compute a *separate* BoVW histogram on each frame. These histograms are concatenated into a high-dimensional representation. Pairwise similarity scores between videos are computed using *global* alignment. In contrast, we employ state-of-the-art features, use a sliding window and select the *most dominant* visual words occurring in each interval. Videos are represented as sequences of these visual words. We employ pairwise local alignment finding similar regions of dominant visual word sequences, while considering metrics on the underlying feature space. Finally, we conduct supplementary experiments on complex datasets [7]. Our experiments show that despite its low dimensionality compared to BoVW, SoDVW retains its discriminative power.

3. METHODOLOGY

Our implementation relies on the state-of-the-art dense trajectories [8] and descriptors (Trajectory, HOG, HOF and MBH) that capture shape and motion. We employ a sliding window and encode temporal information as illustrated next (see Fig. 2): 1. For each window label trajectories by assigning them to their closest codebook centroid-visual word (VW). 2. Collect these VW at each window into sets, *temporal bundles*. 3. Retain the most dominant, in terms of occurrence frequency VWs in each temporal bundle. 4. Order retained VWs within each bundle, based on the average relative temporal location of the trajectories assigned to them. 5. Concatenate the resulting subsequences from each window into a final temporally ordered sequence of dominant visual words (SoDVW).

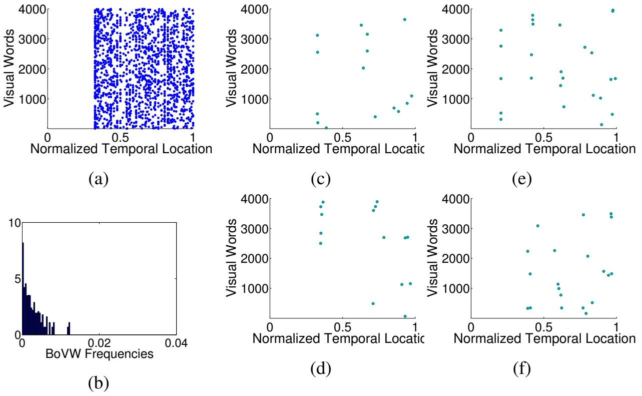

Fig. 3. (a) Visual words across time *b*efore retaining the dominant ones, in terms of BoVW frequencies in (b). Retained visual words distributions for two instances of Stand (c,d) and Sit actions (e,f).

3.1. Temporal Sequence of Dominant Visual Words

Let $\{\mathbf{x}_n\}$ be the set of trajectories of length L extracted from a video and $\mathcal{D}$ the dictionary of K visual words $w_1, \dots, w_K$. Each video is processed using a non-overlapping window of 15 frames. Trajectories which have at least $\lceil \frac{L}{2} \rceil$ overlap with a window are assigned to it. Each window is represented by a set of VW labels occurring within the corresponding time interval. We also retain the relative temporal location $tloc$ of every VW, defined as the mean temporal location of all trajectories assigned to this specific VW.

Each window $i = 1 \dots T$ is represented by a *temporal bundle*, i.e. a set of VW labels. Consequently, each video is represented by a collection of temporal bundles:

$$VWSet_i = \{w_j \,|\, f_i(w_j) \neq 0\}\,, \; i = 1, \dots T, \quad j = 1, \dots, K$$

where $f_i(w_j)$ is the j-th element of local BoVW vector.

Figure 3 depicts VWs occurring within several videos. However, only a few of the VWs occurring in a window are dominant, in terms of occurrence frequency. Therefore, we compute a visual words' occurrence frequency histogram within each temporal window, to retain only the N_d most frequent VWs in the temporal bundle. The resulting temporal evolution of $N_d = 5$ VWs illustrated in Fig. 3 reveals well-formed patterns, which are used as additional information. We further process temporal bundles by sorting the retained VWs according to their relative temporal location $tloc$. In this way, we add temporal information within each temporal bundle, creating subsequences of VWs: $VWSeq_i = [w_{i1}, w_{i2}, \dots, w_{iN_d}]$, where $tloc_i(w_{i1}) \leq tloc_i(w_{i2}) \leq \dots \leq tloc_i(w_{iN_d}) \quad i = 1 \dots T$. By concatenating these VWs subsequences we get a temporally ordered sequence of dominant ones: $SoDVW = [VWSeq_1, VWSeq_2, \dots, VWSeq_T]$.

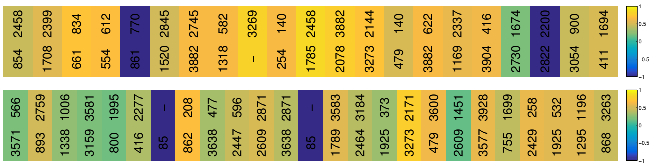

Fig. 4. Pairwise local alignment of dominant visual word sequences. Upper: aligned sequences belonging to the same class; bottom: to different classes. Vertical color strips encode the similarities of aligned visual words. Top: high values (yellow in the colorbar) correspond to visual words that are close. Bottom: much less similarity, i.e. mid ranges of values (green in the colorbar).

3.2. Local Alignment of Visual Action Subsequences

To use the above representation in a SVM, we define the distance between two SoDVWs. We propose determining the similarity between two sequences using the Smith-Waterman (S-W) local sequence alignment algorithm [34]. This has the property of finding the region of highest similarity between sequences, by comparing possible sub-segments and determining the optimal score. Thus, it detects similarities which often are highly divergent due to variations in action execution, duration, and occlusions. This measure is also robust to action-location-within-the-video variability, resulting from loose trimming of original videos to video segments containing a single action. Sequence regions with high dissimilarity are ignored and do not affect the final similarity score.

Given an alphabet Σ, of the K VWs of the dictionary, and two sequences $A : w_{i1}w_{i2}\dots w_{im}$ and $B : w_{j1}w_{j2}\dots w_{jn}$, the algorithm returns the score of their *l*ocal alignment. Additional parameters are the gap penalty p and the similarity matrix S, whose (i, j)-th entry indicates the similarity between the symbols w_i, w_j. Herein we propose a similarity matrix which captures the correlation between VWs:

$$S(w_i, w_j) = -2 * \frac{d(w_i, w_j)}{\max\limits_{k,l=1\dots K} d(w_k, w_l)} + 1,$$

where $d(w_i, w_j)$ is the Euclidean distance.

Next, the S-W algorithm builds a matrix H, where $H(i, j)$ is the maximum similarity score of two segments of the input sequences. For a detailed description, see [34]. The similarity between two sequences is the maximum element of this matrix. Each score is further normalized so that the similarity lies within $[0, 1]$, allowing us to define the distance $D(SoDVW_1, SoDVW_2) = 1 - Similarity(SoDVW_1, SoDVW_2)$. See an example of aligned sequences in Fig. 4.

3.3. Fusion of Temporal Sequence Similarities

For classifying a video via its SoDVW representation using SVMs, we define the kernel: $K(Seq_1, Seq_2) = e^{-\frac{D(Seq_1,Seq_2)}{A}}$, where A is the average pairwise distance between videos. For the BoVW and SoDVW we employ two different measures of similarity i.e. different kernels. The integration with different base kernels is achieved by computing a linear combination of kernels (LCK) within a single SVM. In this work, we experiment with different positive, weight vectors θ for the combination of SW-kernel (K_1) and RBF-χ^2 (K_2) kernel for SoDVW and BoVW representations, respectively: $K = \theta_1 * K_1 + \theta_2 * K_2$.

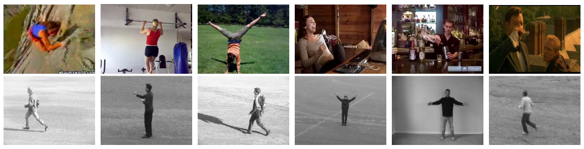

Fig. 5. Sample frames from HMDB51 and KTH datasets.

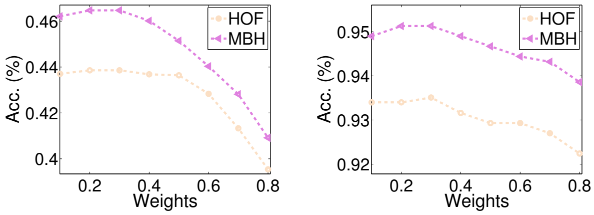

Fig. 6. Performance on (a) HMDB51 and (b) KTH by varying the weight θ_1 parameter of our approach.

4. DATASETS, EXPERIMENTS AND RESULTS

We conduct experiments on the KTH, Hollywood2 and HMDB51 datasets, keeping the original configurations [7, 9, 10] (Fig.5) [1]. We extract improved dense trajectories, using the bounding boxes provided by the authors. The trajectory length is $L = 15$ and we obtain the Trajectory, HOG, HOF, and MBH descriptors. A separate codebook is built per descriptor by clustering 100000 random training samples to $K = 4000$ centers using K-means. For action classification, we use a discriminative SVM classifier [35]. For BoVW we employ the χ^2 kernel, while in the SoDVW case we compute the pair-wise similarity between videos as in 3.2. For multi-class classification we use a "one-vs-all" approach, selecting the higher score class. Multiple descriptors are combined by summing their kernel matrices, while SoDVW and BoVW encoding methods are fused with linear kernel combination (sec. 3.3).

4.1. Experimental Results

Results are presented in Table 1. The proposed SoDVW representation achieves an average accuracy of 38.39% on the challenging HMDB51 dataset using solely temporal information as dominant visual word sequences. These have much smaller length than the BoVW vectors, which in our experimental framework have a constant size of 4000: e.g. the maximum lengths of SoDVW sequences for the training/testing videos of the 1st HMDB51 split are 474 and 350 respectively, with a median value of only 50. Therefore, action recognition results demonstrate the rich information captured in these small sequences and their discriminating power.

The fusion of the two approaches yields better results than the baseline. Improvements on KTH are up to 1%. On HMDB51 dataset, improvements from the baseline are greater ranging from 2% to 5%, leading to 54.05% accuracy. Similar improvement can be observed for Hollywood2. These demonstrate the complementary nature of our approach, which integrates temporal information into the BoVW framework. A side point to stress is

[1] We report average classification accuracy over the three splits for HMDB51 and mean average precision (mAP) on Hollywood2. We use the following default parameters for SoDVW representation, unless otherwise specified: a temporal window of 15 frames, $N_d = 10$ dominant visual words, Smith-Waterman gap penalty $p = 0.1$ and LCK weight vector $[0.3, 0.7]$.

	Method	Traj.	HOG	HOF	MBH	Comb.
KTH	BoVW	90.85	86.67	93.4	94.67	94.09
	SoDVW	83.78	80.76	86.33	87.83	87.83
	BoVW+SoDVW	91.19	86.79	93.51	**95.13**	94.67
HMDB51	BoVW	33.47	29.13	41.26	43.55	52.16
	SoDVW	23.75	18.84	30.61	25.53	38.39
	BoVW+SoDVW	38.32	34.18	43.86	46.47	**54.05**
HOHA2	BoVW	49.82	40.66	52.15	55.35	59.85
	SoDVW	34.3	32.62	40.33	38.91	44.92
	BoVW+SoDVW	53.72	45.56	53.82	57.15	61.34

Table 1. Average recognition accuracy results on KTH and HMDB51 and mean average precision (mAP) on Hollywood2.

set	SoDVW	BoVW	R_1	SoDVW+BoVW	R_2
train	214840	14280000	66.47	14494840	1.015
test	88862	6120000	68.87	6208862	1.015

Table 2. Data employed in SoDVW vs BoVW for HMDB51, measured as the total number of elements contained in the corresponding video representations. In our case we employ 66 to 68 times less data, as shown by the ratio R_1 between the SoDVW/BoVW quantities. We also show the R_2 ratio between the BoVW and the SoDVW+BoVW showing that the increase is negligible.

the amount of data used to compute the SVMs' kernels. Our approach stores much less data per video; that is a vector containing $N_d \times (number\ of\ windows)$ elements, whereas BoVW stores a K-dimensional histogram per video. Providing more data towards this, we pose the following question: What percentage of the total accuracy can our representation achieve using only the most dominant visual words? The answer is quite revealing: as shown in Table 3 this is at least 71% for HMDB51. Thus, using 66 times less data, we reach a significant percentage of the total accuracy.

	HMDB51	KTH	Hollywood2
p_1	73.6	93.35	75.05
p_2	71.03	92.78	73.23

Table 3. Percentage of the total accuracy that SoDVW achieves compared to BoVW (p_1) and SoDVW+BoVW (p_2).

Table 4 shows the effect of the gap penalty. By retaining $N_d = 5$ dominant visual words at each temporal bundle, we experiment with two gap penalties, $p = 0.1$ and $p = 2$ on KTH. A value of 0.1 leads to better performance. This is attributed to the more successful local sequence alignment achieved with this penalty. Setting it to 2, an alignment without gaps, i.e. insertions or deletions, entails a lower cost in comparison to alignments with high-cost gaps. Another parameter is the weight θ_1 that weights the sequence similarity kernel (see Fig. 6). In almost all cases there is an increase to the fused result. By altering the parameter we adjust the relative contribution of our approach, which is maximized at $\theta_1 = 0.3$, and is later on employed based on its consistent performance.

	Traj.	HOG	HOF	MBH
SoDVW ($p = 2$)	82.97	76.94	82.97	85.05
SoDVW ($p = 0.1$)	84.24	79.84	86.67	89.34

Table 4. Action recognition results with varying penalty parameter on the KTH dataset, $N_d = 5$.

Work	Method Note	Year	KTH	HMDB51	Hollywood2
[5]		2014	97.2	24.5	-
[25]		2013	89.7	-	-
[24]	FV	2014	-	65	-
[26]		2008	86.8	-	-
[39]		2008	92.0	-	-
[36]	CD	2014	96.5	53.4	-
[36]	CD+FV	2014	-	58.7	-
[18]	SVM Fusion	2014	-	59.4	-
[16]	FV	2013	-	54.8	63.3
[16]	w/o FV	2013	-	-	58.1
[40]	iDT+FV	2015	-	63.7	-
[4]	DT	2011	94.2	-	58.2
[8]	iDT+BoVW	2013	-	52.1	62.2
[8]	iDT+FV	2013	-	57.2	64.3
[38]	VLAD	2013	-	52.3	62.5
[38]	w/o VLAD	2013	-	45.6	58.5
[37]		2013	98.2	26.9	-
[41]		2014	-	47.6	-
[42]		2013	-	37.3	-
[43]		2012	-	-	60
[44]		2011	93.9	-	53.3
[45]		2014	-	-	59.6
Ours			95.1	54.1	61.3

Table 5. Comparison to temporal-related approaches (upper part), and other methods from the recent state-of-the-art (lower part).

Finally, we compare with other methods reported in the literature. In many cases, as shown in the upper part of Table 5, our method performs better compared with other approaches explicitly modelling temporal information. As far as the rest of the included approaches are concerned, we demonstrate increased performance compared to the trajectory-based approaches, such as the improved/dense trajectories [8], the causality descriptor [36], the action-bank on HMDB51 [37], and Jain *et al.* [38]. The majority of approaches that outperform our method employ Fisher vector, though we reach comparable performance to [16]. Incorporating FV within the proposed framework, which is directly applicable, we expect improved performance due to the complementary information provided by our method. In any case, our approach achieves results within the state-of-the-art using much lower dimensional data.

5. CONCLUSIONS

We introduced a simple yet effective approach that models the temporal structure of actions, representing them as sequences of dominant visual words and measuring their similarity using local sequence alignment. We have demonstrated that combining our representation with BoVW improves recognition performance. Therefore, our representation carries complementary information that captures temporal aspects of actions. In the future we plan to integrate it with other top-performing video representations, such as the Fisher vector, experiment on supplementary datasets and explore the effect of alternative alignment algorithms.

6. REFERENCES

[1] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*. IEEE, 2006, vol. 2, pp. 1447–1454.

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

[3] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*. IEEE, 2009, pp. 2929–2936.

[4] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*. IEEE, 2011, pp. 3169–3176.

[5] P. Agustí, V. J. Traver, and F. Pla, "Bag-of-words with aggregated temporal pairwise word co-occurrence for human action recognition," *Pattern Recognition Letters*, vol. 49, pp. 224–230, 2014.

[6] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," in *European Conference on Computer Vision (ECCV 2010)*, pp. 508–521. Springer, 2010.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *IEEE International Conference on Computer Vision (ICCV 2011)*. IEEE, 2011, pp. 2556–2563.

[8] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision (ICCV 2013)*. IEEE, 2013, pp. 3551–3558.

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

[10] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[11] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*. IEEE, 2008, pp. 1–8.

[13] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, Sept. 2005.

[14] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010)*. IEEE, 2010, pp. 3304–3311.

[15] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision (ECCV 2010)*, pp. 143–156. Springer, 2010.

[16] D. Oneata, J. Verbeek, and C. Schmid, "Action and Event Recognition with Fisher Vectors on a Compact Feature Set," in *IEEE International Conference on Computer Vision (ICCV 2013)*, Sydney, Australia, Dec. 2013, pp. 1817–1824, IEEE.

[17] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *arXiv preprint arXiv:1405.4506*, 2014.

[18] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[19] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, June 2014.

[21] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*. IEEE, 2007, pp. 1–8.

[22] O. Ramana Murthy and R. Goecke, "The influence of temporal information on human action recognition with large number of classes," in *International Conference on Digital lmage Computing: Techniques and Applications (DlCTA 2014)*, Nov 2014, pp. 1–8.

[23] T. Glaser and L. Zelnik-Manor, "Incorporating temporal context in bag-of-words models," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops 2011)*. IEEE, 2011, pp. 1562–1569.

[24] Z. Lan, X. Li, and A. G. Hauptmann, "Temporal extension of scale pyramid and spatial pyramid matching for action recognition," *arXiv preprint arXiv:1408.7071*, 2014.

[25] G. Cheng, Y. Wan, W. Santiteerakul, S. Tang, and B. P. Buckles, "Action recognition with temporal relationships," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2013)*. IEEE, 2013, pp. 671–675.

[26] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *IEEE Workshop on Motion and video Computing (WMVC 2008)*. IEEE, 2008, pp. 1–8.

[27] V. Bettadapura, G. Schindler, T. Plötz, and I. Essa, "Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*. IEEE, 2013, pp. 2619–2626.

[28] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, June 2015.

[29] M. Nagel, T. Mensink, and C. Snoek, "Event fisher vectors: Robust encoding visual diversity of visual streams," in *British Machine Vision Conference (BMVC 2015)*, 2015.

[30] C.-C. Chen and J. Aggarwal, "Modeling human activities as speech," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*. IEEE, 2011, pp. 3425–3432.

[31] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. IEEE, 2014, pp. 780–787.

[32] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell, "A novel sequence representation for unsupervised analysis of human activities," *Artificial Intelligence*, vol. 173, no. 14, pp. 1221–1244, 2009.

[33] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video event classification using string kernels," *Multimedia Tools and Applications*, vol. 48, no. 1, pp. 69–87, 2010.

[34] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[35] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[36] S. Narayan and K. R. Ramakrishnan, "A cause and effect analysis of motion trajectories for modeling actions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. IEEE, 2014, pp. 2633–2640.

[37] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*. IEEE, 2012, pp. 1234–1241.

[38] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*. IEEE, 2013, pp. 2555–2562.

[39] K. Hatun and P. Duygulu, "Pose sentences: a new representation for action recognition using sequence of pose words," in *19th International Conference on Pattern Recognition (ICPR 2008)*. IEEE, 2008, pp. 1–4.

[40] Z. Lan, X. Li, M. Lin, and A. G. Hauptmann, "Long-short term motion feature for action classification and retrieval," *arXiv preprint arXiv:1502.04132*, 2015.

[41] F. Shi, E. Petriu, and R. Laganiere, "Sampling strategies for real-time action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*. IEEE, 2013, pp. 2595–2602.

[42] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 817–827, 2014.

[43] E. Vig, M. Dorr, and D. Cox, "Space-variant descriptor sampling for action recognition based on saliency and eye movements," in *European Conference on Computer Vision (ECCV 2012)*, pp. 84–97. Springer, 2012.

[44] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*. IEEE, 2011, pp. 3361–3368.

[45] M. Sapienza, F. Cuzzolin, and P. H. Torr, "Feature sampling and partitioning for visual vocabulary generation on large action classification datasets," *arXiv preprint arXiv:1405.7545*, 2014.