An evaluation of bags-of-words and spatio-temporal shapes for action recognition

Teófilo de Campos, Mark Barnard, Krystian Mikolajczyk, Josef Kittler,
Fei Yan, William Christmas and David Windridge
CVSSP, University of Surrey
Guildford, GU2 7XH, UK
http://www.ee.surrey.ac.uk/CVSSP/

Abstract

Bags-of-visual-Words (BoW) and Spatio-Temporal Shapes (STS) are two very popular approaches for action recognition from video. The former (BoW) is an un-structured global representation of videos which is built using a large set of local features. The latter (STS) uses a single feature located on a region of interest (where the actor is) in the video. Despite the popularity of these methods, no comparison between them has been done. Also, given that BoW and STS differ intrinsically in terms of context inclusion and globality/locality of operation, an appropriate evaluation framework has to be designed carefully. This paper compares these two approaches using four different datasets with varied degree of space-time specificity of the actions and varied relevance of the contextual background. We use the same local feature extraction method and the same classifier for both approaches. Further to BoW and STS, we also evaluated novel variations of BoW constrained in time or space. We observe that the STS approach leads to better results in all datasets whose background is of little relevance to action classification.

1. Introduction

Human action recognition has become a very active research field in recent years [9, 22]. One possible approach to this problem consists in analysing the output of a Human Motion Capture (HMC) systems using combinations of HMMs [12]. However marker-less HMC is still a very challenging task in uncontrolled environments and with low resolution [9]. Discriminative methods offer viable alternatives which map low level visual inputs directly into actions through classification. One such with promising results is that of Efros et al. [7], which uses simple maps of quantised optical flow vectors as local motion descriptors.

Following the generic object categorisation methods for static images, as in the PASCAL VOC challenges [8], research has been focused on recognising actions without locating them in space and time. Most of the methods in this category follow the Bag-of-visual-Words (BoW) approach using spatio-temporal features [23, 26, 16, 20, 31]. The BoW approach is also used by Ballan et al. [1], however in this case recognition is performed using string kernels to model temporal structure. Pang et al. [21] use BoW as an initial step to build bags of synonym sets and incorporate class-based information in the metrics.

In BoW, descriptors extracted at numerous locations in space and time are clustered into a number of visual words and the video is represented by a histogram of these words. One of the main drawbacks is that any spatial or temporal relationship between descriptors is discarded. For static images, this problem is addressed by building separate kernels for spatial partitions of the images [3, 30] or structured image representations [18]. These methods lead to richer description of the object of interest in its context. Moving from images to video, the importance of context may diminish in many applications, as the same person or object in the same context, can perform different actions. In particular, if the focus is on instantaneous actions (e.g., hitting the ball when playing tennis) then the importance of global context is almost none. A global description such as BoW may lead to a noisy representation while an object-centric approach may have better discriminative power. Moreover, it is often possible to segment acting objects from the background if a static camera is used or if the background can be tracked.

In this paper we provide an evaluation of action recog-

This preprint has been published at the IEEE Workshop on Applications of Computer Vision (WACV) – Winter Vision Meetings, Kona, Hawaii, Jan 5-6 2011 ©IEEE.

The research leading to this paper received funds from EPSRC grants EP/F069421/1 (ACASVA), EP/F003420/1 (ROCS) and from EU PASCAL2. We thank Alex Kläser for providing the HOG3D feature extraction program. Thanks to Ibrahim Almajai and Aftab Khan for useful discussions and for their help in the annotation process.

nition methods on four datasets[1]. The first one is a novel dataset of actions in tennis games, i.e., all actions occur in the same context and they are well localised both in space and time (e.g. hitting the ball). We also present experiments in the following public datasets, in increasing level of background complexity: Weizmann [11], KTH [25] and UCF sports [24]. In these datasets, actions are defined by video sequences and may consist of cyclic motions (such as walking). In the case of UCF sports, the categories are often well described by the background context.

We evaluate the standard BoW approach, its variant in which features are localised at the acting person, termed Spatially restricted BoW, or SBoW, and a method based on the Spatio-Temporal Shapes (STS) [11], i.e., an action is treated as a single 3D shape in the spatio-temporal block. In all three approaches, we use the same basic local features with different support in space and time. Our results show that the STS approach outperforms BoW by a large margin for the detection of instantaneous actions in our tennis dataset. In the Weizmann and KTH datasets, STS also outperforms BoW and SBoW. In the UCF sports dataset, STS leads to the same performance as BoW, but SBoW outperforms both by a small margin, suggesting that foreground focus is important but the actions are not structured well enough for STS.

The following section details the methods evaluated. Section 3 provides information about the datasets and methods used to extract bounding boxes of acting subjects. Next, experiments are presented in Section 4 and the paper concludes in Section 5.

2. Methods

This section describes the local feature extraction method and the approaches in which features are combined in order to classify video sequences.

2.1. Local feature descriptor and the STS method

The most popular spatio-temporal feature descriptors are three dimensional generalisations of SIFT [17] or local histograms of oriented gradients (HOG) [6]. They range from methods which simply compute the 2D SIFT and concatenate the descriptor with local motion vectors [27] to methods which actually compute multi-scale oriented gradient histograms in 3D blocks [26, 14]. The latter [14], dubbed HOG3D, uses polyhedral structures for quantisation of the 3-D spatio-temporal edge orientations to avoid the singularities in the use of polar coordinate systems (as done in [26]). Another advantage of HOG3D [14] is its computational efficiency due to the use of three-dimensional integral images. In the benchmark experiments of [31], HOG3D has proven to be among the state-of-the-art methods for the BoW approach. For these reasons, we chose this method as the local spatio-temporal descriptor in all our experiments.

For a given spatio-temporal block, HOG3D splits it into $M \times M \times N$ sub-regions (M for spatial and N for temporal splits). Within each region, this method counts how many 3D gradients are roughly aligned with each of the directions of a polyhedral structure. This computation is done efficiently using basic geometric operations. In [14] the authors claim that, for BoW, the best performance in the validation set of the KTH dataset was obtained with an icosahedron (i.e., 20 orientations), $M = 4$ and $N = 3$, giving a total of 960 dimensions. This may seem large, but the dimensions of the obtained feature vector are little correlated. In our preliminary experiments, we found that the discriminative power of this descriptor is reduced if less than 500 dimensions are used after PCA. We therefore do not apply any dimensionality reduction.

The temporal and spatial support of such descriptors were also optimised in [14], using the validation set of KTH. We found experimentally that a larger temporal support of 12 frames gives better performance for the STS method described in this section.

The spatio-temporal local descriptors can be extracted at densely distributed locations [31], but to improve their efficiency, spatio-temporal keypoint detectors (e.g. [16, 20]) or even random selection of locations [26] has been used. In a number of application domains, such as surveillance and sports (e.g. football in [7]), the background can be tracked and used for foreground segmentation to extract features.

Gorelick et al. [11] proposed to model actions as space-time shapes (STS) by describing the spatio-temporal block where the action is located as a 3D binary shape. The recognition is then approached as a matching problem. To build the descriptor, binary human silhouettes are extracted from a video and grouped as space-time 3D shapes. A 3D shape analysis method is proposed and leads to excellent results on a dataset with uniform and static background. The method has been shown to be robust to some level of deformation and partial occlusion of the silhouettes. However, more challenging data with moving background may lead to highly fragmented silhouettes (e.g. blobs shown at the bottom of Figure 2), which would decrease the performance of that method. In contrast to STS, the HOG3D uses greyscale images rather than binary, thus it does not require pixel-level person segmentation. If a bounding box is given, including all relevant foreground blobs of the acting object, HOG3D is not affected by fragmentation. In this paper we therefore evaluate HOG3D as a descriptor for STS-based action matching. An additional reason for this choice of descriptor is that it has previously been evaluated with BoW methods [31], which provides a common ground for com-

[1] A set of benchmark experiments is presented in [31], but they differ from ours as they focus on different feature extraction methods (all using BoW), whereas we focus on the representation methods, e.g. BoW vs STS.

parison.

In our STS experiments, a single HOG3D descriptor is extracted for each detected actor at the time instance in which the action is classified. The extracted 960D vector is then passed directly to a classifier, without an intermediate representation. For problems in which the aim is to classify the activity in a video sequence rather than an instantaneous action, we use STS at a number of temporal windows within a video sequence. The classification results are then combined using a voting scheme.

2.2. BoW-based methods

We investigate the original bags-of-spatio-temporal-words (BoW) method and two novel variations: the spatially-constrained BoW (SBoW) and the local BoW (LBoW) methods.

All descriptors in the training set are clustered using the k-means algorithm into $|V| = 4000$ clusters, following Kläser et al. [14]. In our preliminary experiments with a number of values of $|V|$, we observed an asymptotic growth in performance up to $|V| = 4000$ which hints that this size does not over-fit to our training sets.

We used a hierarchical k-means process, first the data is clustered into 40 high level clusters and then 100 lower level clusters. A histogram is then produced for each frame of the videos in the training set. The 4000 bin histogram is populated using two techniques: hard and soft voting. Hard voting is the standard vector quantisation method used in BoW. Soft voting uses the codeword uncertainty method presented in [29] where the histogram entry of each visual codeword w is given by

$$UNC(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{K_\sigma(D(w,r_i))}{\sum_{j=1}^{|V|}K_\sigma(D(w_j,r_i))},$$

where n is the number of descriptors in the image, $D(w,r_i)$ is the Euclidean distance between codeword w and the descriptor r_i, K is a Gaussian kernel with smoothing factor σ and V is the visual vocabulary containing the codeword W.

In the initial presentation of this method the authors estimated the value of the smoothing factor σ experimentally using a training and validation set. In our case we estimated σ directly from the data by taking one standard deviation of the distribution of distances from descriptors to their cluster centres. This method proved to be much faster while still producing a reasonable estimate of σ. The *Codeword Uncertainty* method of histogram generation has been shown to perform well in the PASCAL Visual Object Classification challenge [28].

We also follow insight from the commonly used spatial pyramid kernels [5] and evaluate a set of spatio-temporal kernels by dividing the spatio-temporal block of each acting segment in the following configurations in RxCxT (R and C splits in space, and T splits in time): 1x1x1, 2x2x1, 3x1x1, 1x3x1, 1x1x3, 2x2x3, 3x1x3 and 1x3x3. For each cell, its descriptors are accumulated, generating a 4000D histogram. For each of these configurations, the histograms are concatenated, generating a $4000 \times R \times C \times T$ dimensional vector.

2.2.1 Spatially-constrained BoW (SBoW)

In order to investigate the importance of foreground and context, we propose to use bounding boxes of located actors to restrict feature extraction. Dense sampling is used and only features whose centre is within bounding boxes are considered when building BoW histograms. In the SBoW experiments, a single histogram is built for each video, in the same way as with BoW. The bounding boxes are obtained in the same way as for the STS method, further detailed in Section 3.

2.2.2 Local-BoW (LBoW)

A combination of spatial and temporal constraints is also explored, i.e., BoW histograms are built for each temporal window, spatially restricted by the actor's bounding box. Classification is done per temporal window, in a similar way to STS. In our LBoW experiments, the HOG3D descriptors are extracted densely within the spatio-temporal block of the located actor. The spatial support is set to the following range of scales of the actor's bounding box: $w_s = h_s = \mathcal{B} \cdot \{1,2,3,4\}/3$. Descriptors are sampled at 5 instances of time, 4 scales and *up to* 9×9 positions per frame. Features with larger scale and further away from the centre of the block are sampled less densely, resulting in 934 vectors per bounding box.

2.3. Classification

We employ kernel Fisher discriminant analysis (kernel FDA) [19], which has lead to better results than SVM in [33]. We adopt a spectral regression based implementation of kernel FDA [4], which avoids the expensive eigen decomposition. As a result, it is much more efficient than both standard kernel FDA implementations and SVM.

Although kernel FDA can be implemented as a multi-class classifier, we obtained better results by splitting the C-class problems into C pairwise classification problems using a one-against-all scheme and combining the results using the maximum a posteriori among the C classifiers.

Since both the HOG3D descriptor and the BoW representations are based on histograms, one of the most appropriate kernels is the RBF with χ^2 statistics: $K(\mathbf{x},\hat{\mathbf{x}}) = \exp[-\frac{1}{\sigma}D(\mathbf{x},\hat{\mathbf{x}})]$, where $D(\mathbf{x},\hat{\mathbf{x}}) = \frac{1}{2}\sum_{k=1}^{K}\frac{[x_k-\hat{x}_k]^2}{x_k+\hat{x}_k}$, and k is the index of the histogram bin (i.e. the dimension of the vectors) [2]. Following a frequent approach in image

Footage	length	play shots	serve	hit	non-hit
singles 03	35min	80	76	219	943
doubles 09	30min	34	46	167	1351

Table 1. Statistics of our tennis primitive actions dataset.

categorisation, $\sigma = \frac{1}{N^2}\sum_{i,j}^{N} D(\mathbf{x}_i, \mathbf{x}_j)$, for all $\mathbf{x}_i, \mathbf{x}_j$ in the training set. This kernel has also been used in the BoW-based evaluations of Wang et al. [31].

3. Datasets and Detection Methods

In order to perform a comparison we have selected three of the most commonly used human action recognition databases: KTH [25], Weizmann [11] and the UCF sports database [24]. In addition, we present a novel dataset of instantaneous actions in tennis games. The following sections give further details about each of these datasets.

Person or actor detection is not in the scope of contributions of this paper. Therefore, for the methods that require actors localisation (STS, LBoW and SBoW), we use heuristics to detect moving blobs in images.

3.1. Instantaneous actions in tennis

This dataset was built with the goal of evaluating primitive player action recognition in tennis games. The player actions required for automatic indexing of tennis games are *serve* and *hit*. A *hit* is defined by the moment a player hits the ball with a racket, if this is not a *serve* action. A third class, called *non-hit* was also used and it refers to any other action. If a player swings the racket without touching the ball, it is annotated as *non-hit*. No distinction is made between near and far players in our annotation.

We used footage from two TV broadcasts of tennis games to build this dataset. Both are matches of females in the Australian Open championships. For training, we used the final game of singles from the 2003 championship, which has a bright green court (see Figure 1-top). For testing, we used the final game of doubles of 2009 (see Figure 1-bottom). Table 1 gives some statistics of this dataset. Both broadcasts include close-ups and commercial breaks as well as valid game shots (dubbed *play shots*). Shot boundaries were detected using colour histogram intersection between adjacent frames. Each shot is then classified as *play shot* or *break* using a combination of colour histogram mode and corner point continuity. False positives are then pruned by a tennis court detection method as detailed in [13]. The number of detected play shots in each video sequence is shown in Table 1.

This is a relatively small dataset, but it is quite challenging, with high levels of motion blur and varying player's size between 30 and 150 pixels. There is a large variation in the number of training and test data for different categories. In order to evaluate the classification, we compute the area

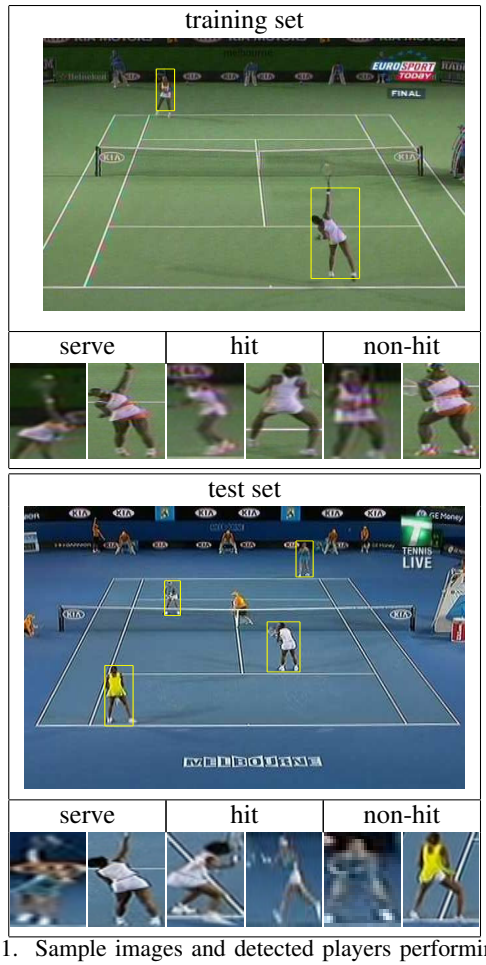

Figure 1. Sample images and detected players performing each action from our dataset of tennis primitive actions.

under the ROC curve (AUC) for each class and average the result, obtaining the mean AUC (mAUC).

In tennis games the background can be tracked reliably, which makes it possible to robustly segment player candidates, as explained in [13]. We extract bounding boxes of the moving blobs and merge overlapping the ones. Next, geometric and motion constraints are applied to further remove false positives. A map of likely player positions is built by cumulating player bounding boxes from the training set. A low threshold on this map rejects bounding boxes from the umpires and ball boys/girls. Figure 1 show some resulting players detected in this manner, performing different actions.

The above algorithm gives player locations in space. To detect the time of action events, we apply the tennis ball tracker of [32]. This method uses a multi-layered data association scheme with graph-theoretic formulation for tracking objects that undergo switching dynamics in clutter. The points at which the ball changes its motion abruptly correspond to key events such as hit and bounce. The generalised edge-preserving signal smoothing is used to detect

these motion changes.

3.2. Weizmann and KTH action datasets

The Weizmann [11] and KTH [25] datasets contain videos of a single person performing actions with uncluttered backgrounds. Therefore the spatial localisation of the action is reliable, but actions are not instantaneous. Instead, they are annotated as video sequences, without detailed temporal delimitation.

In the case of Weizmann, both camera and background are static, and the background image is provided. Therefore person detection is trivial by background subtraction, and the obtained binary maps give reliable bounding boxes. This dataset contains 9 people performing 10 actions in a total of 93 relatively short videos. The evaluation protocol is a leave-one-person-out and the results are normally presented as mean and standard deviation of the accuracy.

The KTH dataset contains 6 actions, 25 subjects, 4 settings: outdoors, with scale variations, with different clothes and indoors. There is a total of 2391 video samples, each annotated as 4 or 5 action sequences which are treated as individual samples. The sequences are relatively long. This dataset has an evaluation protocol defined, with 16 subjects for training and validation, and 9 subjects for testing.

In contrast to Weizmann, the cameras used to collect the KTH dataset are of low quality, with automatic gamma correction and a high level of noise. The outdoors sequences were captured by a hand-held camera, so there is motion in the background in most of the videos. In the indoor videos, people's shadows are cast on the wall behind them. All these factors, as well as the greyscale format used means that person segmentation presents some degree of challenge.

We use a combination of two methods to detect bounding boxes: a smoothed motion map and a pixel classification method. The motion map is computed by subtracting consecutive images, thresholding and filtering the output. The pixel classification method uses pixels near the image border to model the background of each frame and a threshold is applied on the distance from this model. Both maps are combined by the AND operator which gives satisfactory results. Some examples are shown in Figure 2.

3.3. The UCF sport disciplines dataset

The UCF sports database [24] has short videos of different sport disciplines obtained from TV broadcasts. Due to copyright issues, the complete set of classes described in [24] is not available publicly. We follow the subset used in [31], which contains 10 disciplines: diving, golf swinging, kicking, weight lifting, riding horse, running, skateboarding, swinging on the pommel horse or on the floor, swing around the high bar and walking. In total, 150 video sequences are available. As in [31], we expand the training

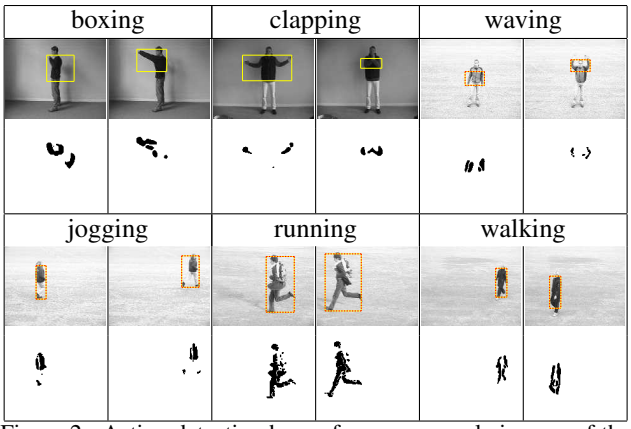

Figure 2. Action detection boxes for some sample images of the KTH dataset (upper rows) and the obtained masks used to extract the bounding boxes (bottom rows).

Figure 3. Sample bounding box crops of stills of the UCF dataset.

set by using mirrored videos. The evaluation is done in a leave-one-video-out setup resulting in 150 train/test experiments. Obviously, the mirrored version of the test video is not included in the training set.

Bounding boxes of the acting people are available for each frame of this dataset. As shown in Figure 3 the bounding boxes do not always include key discriminative elements of the action. For instance, the golf club is not always visible and only a small portion of horses appear in the boxes of 'riding horse' action samples. The same happens with skating. Occlusions may also happen as in the sample for running.

4. Experiments and Results

Our experiments investigate the role of object localisation, both in space and in time, for action classification in video. We present them for each dataset evaluated.

4.1. Classification of instantaneous actions in tennis

In the dataset of Section 3.1 multiple actions occur in the same frame (e.g. non-hit and hit) and the actions occur at different instants of the same video sequence. It is

temporal split	spatial split 1x1	1x3	2x2	3x1	MK
x1	78.5	78.2	79.6	79.5	80.6
x3	84.4	82.3	82.8	84.4	**84.5**

Table 2. Results with the Tennis actions dataset – mean AUC (%) obtained with LBoW using different spatio-temporal pyramid kernels and their combinations. The STS single feature method resulted in mean AUC of **90.3**%.

	non-hit	hit	serve
non-hit	1068	182	117
hit	36	119	14
serve	2	3	41

Table 3. Results with the Tennis actions dataset – confusion matrix of the best method, STS, for thresholds selected so that the true positive rate is 77.62% and the false positive rate is 22.38%.

	STS	LBoW	[11]
mean	94.43	86.50	97.83

Table 4. Results on the Weizmann dataset – mean accuracy (in %) *per temporal window*. A window was sampled at each frame of the video sequences and classified individually. LBoW was computed with features sampled densely at each spatio-temporal location.

therefore not possible to process a whole sequence to build a single BoW histogram. For this reason, only the STS and the Local BoW (LBoW) representations were evaluated on a per-frame basis.

Table 2 shows the results with the spatio-temporal kernels used for LBoW. MK stands for multiple kernel combination, which is an average of kernels. MKx1 and MKx3 lead to marginal improvements over the individual kernels. We therefore do not show experiments with single kernels for the other datasets. The best individual kernels were 1x1x3 and 3x1x3, both with mAUC of 84.4%, while MKx3 gave mAUC of 84.5%. In the same dataset, STS gave a mAUC of 90.3%. The ROC curves obtained with a single descriptor (STS) and with MKx3 are in Figure 4. Table 3 shows a confusion matrix for STS.

4.2. Weizmann actions

Tables 4 and 5 shows that the single spatio-temporal descriptors approach (STS) outperforms BoW and LBoW and gives state-of-the-art results. Our HOG3D-based STS method was outperformed by Gorelick et al.'s STS method [11], which is a discriminative descriptor of a sequence of binary silhouettes, i.e., it relies on the quality of the silhouettes. The HOG3D is a generic spatio-temporal descriptor originally proposed within a BoW framework and with parameters optimised for that use. The fact that its result is comparable with Gorelick's highlight the richness of the HOG3D representation.

The LBoW method performed very weakly in this

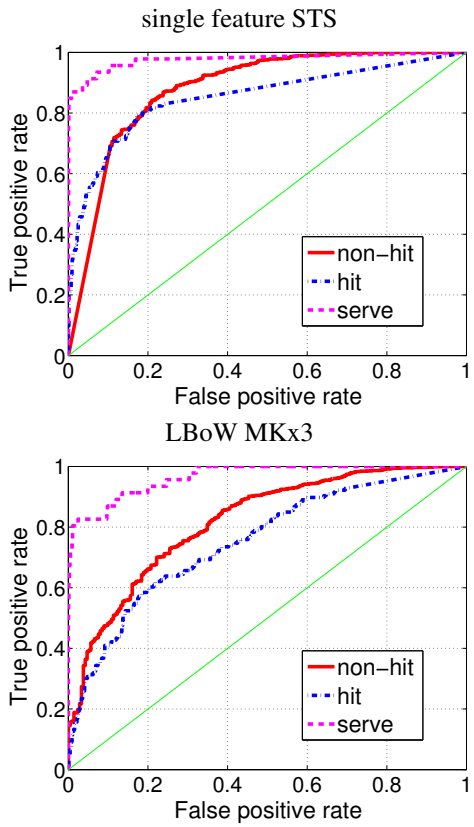

Figure 4. ROC curves obtained with STS (top) and with LBoW with MK combination (bottom). The obtained mean AUC are 90.3% and 84.5%, respectively.

	STS	SBoW	BoW hard	BoW soft	[14]
mean	**96.67**	85.00	86.11	90.00	84.3

Table 5. Results on the Weizmann dataset – accuracy (in %) *per video sequence* BoW was computed using HOG3D's features extracted at locations detected by Laptev's spatio-temporal keypoint detector [16].

dataset, giving results that are almost ten percent lower than the current state of the art. This is rather disappointing for a method that uses a large number of features extracted at each time window. For this reason, we do not present the results with LBoW on the next datasets. In all our BoW-based experiments in this dataset and in the following sections, various pyramid kernels did not lead to significant difference in the results, so we show results with 1x1x1 kernels.

4.3. KTH actions

Table 6 shows the results obtained on the KTH actions dataset. Again, single feature STS gave results that are better than BoW methods. Only the method of [10], based on features data mining, outperformed the STS.

STS	SBoW	BoW hard	BoW soft	[14]	[31]	[24]	[10]
93.52	79.51	88.00	90.00	91.4	92.10	88.66	95.50

Table 6. Results (accuracy in %) on the KTH dataset, per video. For the STS method, a window was sampled at every 6 frames of the video sequences and classified individually. This gave an average detection accuracy of 82.52% per temporal window. The combination with the voting scheme gave 93.52% (shown above). BoW was computed using HOG3D's features extracted at locations detected by Laptev's spatio-temporal keypoint detector [16]. In the [31] column, we report the best result of [31]: HOF with Haris3D detector.

STS	SBoW	BoW hard	BoW soft
80.00	83.33	81.38	80.80

Table 7. Results per video sequence (in % of accuracy) on the UCF sports dataset. The BoW-based methods used dense feature extraction, because Wang et al. [31] have shown that this gives better results than keypoint-based methods in this dataset. For STS, the mean accuracy per individual frame was of 77.64 ± 37.34.

	dive	golf	kick	weight	horse	run	skate	pommel	bar	walk
diving	14	0	0	0	0	0	0	0	0	0
golf	0	14	0	0	0	0	0	1	0	3
kick	1	1	18	0	0	0	0	0	0	1
lift	0	0	0	5	0	0	0	0	0	1
ride	0	0	0	0	12	0	0	0	0	0
run	2	0	3	0	1	7	0	0	0	0
skate	0	1	0	0	0	0	5	0	0	6
pommel	0	0	0	0	2	0	0	18	0	0
bar	0	0	0	0	0	0	0	0	12	1
walk	0	1	0	0	0	0	0	1	0	20

Table 8. Confusion matrix of the SBoW method with 1x1x1 soft histograms on the UCF sports dataset.

4.4. UCF sport disciplines

Table 7 shows the results for the UCF sports dataset. Notice that in this case the best performing methods are the BoW-based ones which use a single histogram to represent a whole sequence rather than the methods based on classification per time window. This is expected, since the set of classes in this dataset represent different disciplines, thus global descriptors are more discriminative. However, a spatial focus on the foreground region does improve the discrimination, given that SBoW performed better than BoW. Table 8 shows the confusion matrix obtained with SBoW.

5. Concluding remarks

This paper presented a comparative evaluation of different methods to represent action for classification. Using the HOG3D as a common ground method for spatio-temporal feature extraction, we evaluated approaches with varied degree of representation of context. At a more global and stochastic level of representation is the popular Bags-of-visual-Words (BoW) approach, which uses numerous local features to build a vectorial representation of a video sequence. At a more local and foreground-focused level is the Spatio-Temporal-Shape (STS) approach, which uses a single feature extraction on a detected bounding box, followed directly by classification, with no intermediate representation. At a conceptually intermediate level, we also proposed a variation of BoW with Spatial restriction to the actors bounding box (SBoW). Like BoW, SBoW gives a global representation built per video sequence. Additionally, we proposed a local representation (LBoW) which gives one representation restricted in space and time. For datasets in which each action is represented by relatively long video sequences, this method works by classifying all the time windows and then combining the results with a voting scheme. The same applies for STS. For the BoW-based representations, we evaluated spatio-temporal pyramid kernels (RxCxT, with divisions in rows, columns and time, respectively).

Our experiments were done on four datasets with increasing level of background complexity: a novel dataset of tennis actions, the Weizmann, KTH and UCF sports datasets. In all cases, except for the UCF sports dataset, STS outperformed all the variations of BoW. This showed that, given that the action is localised, even a single local descriptor per video can often lead to better results than BoW-based methods which extract features throughout the sequence. In the UCF sports dataset, the spatially restricted BoW (SBoW) outperformed both global BoW and STS. This shows that the focus on the foreground was helpful, but action is better represented as unstructured sets of local features.

For future work, we suggest that further investigation should be done in order to automatically learn the trade-off between context and foreground. Further assessment shall be done with other feature extraction techniques that are complementary to HOG3D as well as other datasets such as the Hollywood-Localisation dataset [15]. Another possible research direction is a decomposition of actions into primitive actions, i.e., instantaneous and local elements of action. The STS approach seems appropriate for primitive action detection.

References

[1] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra. Video event classification using string kernels. In *Multimedia Tools and Applications*, volume 48, pages 69–87, 2010.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, April 2002.

[3] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proc of the International Conference on Image and Video Retrieval*, 2007.

[4] D. Cai, X. He, and J. Han. Efficient kernel discriminant analysis via spectral regression. In *International Conference on Data Mining*, 2007.

[5] J. Choi, W. J. Jeon, and S.-C. Lee. Spatio-temporal pyramid matching for sports videos. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc IEEE Conf on Computer Vision and Pattern Recognition, San Diego CA, June 20-25*, 2005.

[7] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc 9th Int Conf on Computer Vision, Nice, France, Oct 13-16*, pages 726–733, 2003.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge (VOC) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/, 2009.

[9] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2/3):77–254, 2006.

[10] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proc 12th Int Conf on Computer Vision, Kyoto, Japan, Sept 27 - Oct 4*, 2009.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[12] N. Ikisler and D. Forsyth. Searching video for complex activities with finite state models. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition*, June 2007.

[13] J. Kittler, W. J. Christmas, F. Yan, I. Kolonias, and D. Windridge. A memory architecture and contextual reasoning for cognitive vision. In *Proc. Scandinavian Conference on Image Analysis*, pages 343–358, 2005.

[14] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *British Machine Vision Conference*, pages 995–1004, sep 2008.

[15] A. Kläser, M. Marszałek, C. Schmid, and A. Zisserman. Human focused action localization in video. In *International Workshop on Sign, Gesture, Activity*, 2010. in conjunction with ECCV.

[16] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 108:207–229, 2007.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int Journal of Computer Vision*, January 2004.

[18] J. J. McAuley, T. de Campos, G. Csurka, and F. Perronnin. Hierarchical image-region labeling via structured learning. In *Proc 20th British Machine Vision Conf, London, Sept 7-10*, 2009.

[19] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *IEEE Signal Processing Society Workshop: Neural Networks for Signal Processing*, 1999.

[20] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int Journal of Computer Vision*, 2008.

[21] L. Pang, J. Cao, J. Guo, S. Lin, and Y. Song. Bag of spatio-temporal synonym sets for human action recognition. In 16^{th} *International Multimedia Modeling Conference (MMM)*, volume 5916 of *LNCS*, pages 422–432, Chongqing, China, January 6-8 2010. Springer.

[22] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, June 2010.

[23] H. Riemenschneider, M. Donoser, and H. Bischof. Bag of optical flow volumes for image sequence recognition. In *Proc 20th British Machine Vision Conf, London, Sept 7-10*, 2009.

[24] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *Proc IEEE Conf on Computer Vision and Pattern Recognition, Anchorage, AK, June 24-26*, 2008.

[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proc Int Conf on Pattern Recognition (ICPR)*, Cambridge, UK, 2004.

[26] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc of the ACM Multimedia*, Augsburg, Germany, September 23–28 2007.

[27] H. Uemura, S. Ishikawa, and K. Mikolajczyk. Feature tracking and motion compensation for action recognition. In *In British Machine Vision Conference*, 2008.

[28] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, September 2009.

[29] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, July 2010.

[30] V. Viitaniemi and J. Laaksonen. Spatial extensions to bag of visual words. In *ACM International Conference on Image and Video Retrieval (CIVR)*, Santorini, Greece, July 8–10 2009.

[31] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc 20th British Machine Vision Conf, London, Sept 7-10*, 2009.

[32] F. Yan, W. Christmas, and J. Kittler. Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences. *Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1814–1830, 2008.

[33] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler. Lp norm multiple kernel fisher discriminant analysis for object and image categorisation. In *Proc IEEE Conf on Computer Vision and Pattern Recognition, San Francisco, CA, June 15-17*, 2010.