Spatio-temporal Video Representation with Locality-Constrained Linear Coding

Manal Al Ghamdi, Nouf Al Harbi, and Yoshihiko Gotoh

University of Sheffield, UK

Abstract. This paper presents a spatio-temporal coding technique for a video sequence. The framework is based on a space-time extension of scale-invariant feature transform (SIFT) combined with locality-constrained linear coding (LLC). The coding scheme projects each spatio-temporal descriptor into a local coordinate representation produced by max pooling. The extension is evaluated using human action classification tasks. Experiments with the KTH, Weizmann, UCF sports and Hollywood datasets indicate that the approach is able to produce results comparable to the state-of-the-art.

1 Introduction

There exist numerous applications in the field of computer vision, including video information extraction and retrieval, classification, summarisation, surveillance and human-computer interaction to name a few. A number of techniques have been put forward to progress this area, a rough sketch of which is given in the next section. This paper presents a spatio-temporal extension of the locality-constrained linear coding (LLC) scheme for video classification tasks. In order to detect interest feature points, the dense 2D scale-invariant feature transform (SIFT) is replaced with spatio-temporal SIFT (ST-SIFT). The ST-SIFT is able to extract effectively the significant invariant local points in the spatial and the temporal domains.

The approach consists of two principle stages: The first stage involves transformation of a 3D video signal into spatio-temporal pyramids, followed by extraction of the interest features from spatial and spatio-temporal planes using a ST-SIFT detector. In the second stage, the LLC is applied on the extracted descriptor in order to encode the local descriptors with similar basis from a codebook. The approach is evaluated using a human action classification tasks as a benchmark. To this end, KTH, Weizmann, UCF sports and Hollywood datasets are used in the experiment, resulting in performances comparable to the state-of-the-art. The contributions of this work can be summarised as follows:

* Extension of the current LLC scheme from a 2D image to a spatio-temporal video signal;
* Provision of a robust schema to represent a human action signal;
* Application of the spatio-temporal LLC for human action classification achieving the state-of-the-art performance on several benchmarks.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 101–110, 2012.

2 Related Work

There have been a number of visual content based approaches developed in recent years. Harris and Forstner's work on interest point operators and the detection of local structures in space-time was taken up and further developed by Laptev and Lindeberg [1]. The concept of correlating a 2D image to a 3D space-time volume was explored by Shechtman and Irani [2], as this approach facilitates the correlation of dynamic behaviours and actions. Alternatively, the maximisation of mutual information (MMI) was utilised by Lio and Shah [3] in order to pick the best total of words for a bag-of-words algorithm. Identification of natural human behaviour within a variety of different and true-to-life video settings was attempted by Laptev *et al.* [4], while Klaser *et al.* [5] used histograms of oriented 3D spatio-temporal gradients to develop a local descriptor.

The work undertaken by Wong and Cipolla [6] made use of information to derive a set of motion recognition interest points; the spatio-temporal interest points with equal time scale-invariant (spatial and temporal) are discussed by Willems *et al.* [7]. SIFT has been successful in various image processing applications for locally detecting and describing interest points [8]. SIFT extension can be categorised into three groups: (1) extension of the descriptor part only, combined with 2D detectors provided by Scovanner *et al.* [9], (2) a full 3D spatial extension such as n-SIFT by Cheung and Hamarneh [10], and (3) a combination of different approaches to separately describing motion and appearance as MoSIFT by Chen and Hauptmann [11].

Despite the fact that a variety of algorithms have been suggested to complete the task, several problems remain; the notable one is that, in general, a feature-point method tends to sparsity, with the implication of high complexity levels for the selection of the model and learning. In response to this, a number of approaches were proposed that involved combined features learning techniques, dimensionality reduction approaches and clustering methods. Robust sparse coding (RSC) scheme, for instance, was presented by Yang *et al.* [12], where sparse coding (SC) was considered as a sparsity-constrained robust regression problem. RSC improved the performance of the original SC and proved its effectiveness in handling facial occlusions.

Developed by Grauman and Darrell [13], the spatial pyramid matching (SPM) used vector quantisation (VQ) to solve a constrained least square problem. The mechanism of the SPM was based on partitioning the input image into sub-regions; local features of each region were extracted using the appropriate descriptor in order to generate descriptor layers. A codebook of M entries was applied to quantise each descriptor and obtain code layers. To achieve the SPM layer, the product of all sub-regions was grouped by averaging and normalising into a histogram. Histograms of all sub-regions were concatenated together in order to generate the final representation of the image. However the SPM method discarded the similarity between similar descriptor; because of the large quantisation error, the VQ code for comparable descriptor could lead to completely different response. More recently Yang *et al.* [14] developed an extension of the SPM by replacing VQ with SC. It was derived by relaxing the cardinality restriction constraint of VQ, so each descriptor could be encoded by multiple bases. Although SC had shown remarkable effectiveness in representing feature quantisation, it suffered from two limitations: (1) even if there was a simple variation in local features, the response of basis in dictionary could be quite different, (2) it eliminated the interdependence and relationships between local features, which adversely affected the image representation.

In order to overcome these limitations Wang *et al.* [15] proposed the LLC, utilising K nearest codewords to encode descriptors within the Euclidean space. In LLC, each descriptor was more precisely encoded by different bases, and LLC code was able to detect the interdependence between similar descriptors by sharing bases. By making use of locality constraints, LLC projected descriptors into their respective local coordinate systems. The representational output was then created by max pooling the coordinates of these projections. In conjunction with a linear classifier, this method was a considerable improvement over the non-linear SPM that was traditionally used, giving the best available performance on a number of standard benchmarks. The LLC approach provided (through the objective function) an analytical solution, as opposed to the computational response generated by SC strategies. It also worked efficiently due to its fast approximation method, which involved initially performing a k-nearest neighbour search, allowing a subsequent small constrained least square fitting problem.

3 Spatio-temporal Coding

This section presents the locality constrained spatio-temporal coding technique that considers the locality of the manifold structure in the input space.

3.1 Spatio-temporal SIFT

The ST-SIFT algorithm is developed to represent video content with invariant interest points. These points contain the amount of information sufficient to describe video streams. Most of the previous studies extended the SIFT algorithm spatially to extract the extrema from 3D images [10,11,16] or detect 2D interest points and describe them with a 3D descriptor [9]. The ST-SIFT, on the other hand, detects the spatially distinctive points with sufficient motion information at multi-scales. To achieve the invariance in both space and time, a spatio-temporal difference of Gaussian (DoG) pyramid is calculated first. The common points between three spatial and temporal planes carry vital information. The ST-SIFT algorithm is outlined below.

Spatio-temporal DoG Pyramid. For a video sequence with the frame size of $W \times H$, let $I(x, y, t)$ denote a pixel at location (x, y) in frame t. We construct the Gaussian pyramid of N levels where N is determined from the frame size. G_i $(i = 0, \dots, N-1)$ represents each level of the pyramid, where the highest level $G_0(t)$ corresponds to the original video frame sequence. This process leads to the multi-level spatio-temporal Gaussian and the DoG pyramids. Incremental convolution of video signal I with the 3D Gaussian filter G results in the scale space L of the first level:

$$L(x, y, t, \sigma, \tau) = G(x, y, t, \sigma, \tau) * I(x, y, t, \sigma, \tau) \tag{1}$$

with multiple scales S separated by a constant value of $K = 2^{1/S}$. Following Lowe [8], $S + 3$ scales are generated to guarantee that local extrema detection will cover the complete octave. To produce a lower level the signal is spatially and temporally downsampled with the Gaussian at scales σ and τ. This yields a level with the lower frame rate and frames of the smaller size. The frame size at level G_i is $W/2^i \times H/2^i$, and $G_i(t)$ matches $G_0(2^i t)$ at time t. The next step is to construct a DoG pyramid; for each level in the Gaussian pyramid, a DoG of one lower octave is derived by subtracting the Gaussian of the adjacent scales.

Interest Points Detection. Once the DoG pyramid is constructed, local extrema of the adjacent scales in the xy, xt and yt planes are compared. Lopes *et al.* presented an approach to forming a spatio-temporal volume by stacking a set of frames from a video signal [17]. There are three directions to slice this volume into planes. One can slice through the spatial axis to create xy planes. Alternatively one can create a sequence of planes from the temporal axis combined with either of x or y spatial axis. Extrema are detected from each slice of the spatio-temporal pyramid separately, and the union of common extrema in three directions are selected as interest points. In the end filtering may be applied to remove noisy points and edges.

3.2 Conventional 2D LLC

The LLC is a coding scheme proposed by Wang *et al.* [15] to project individual descriptors onto their respective local-coordinate systems. Locality is more important than sparsity with LLC because, although locality implies sparsity, the reverse does not hold. The use of the locality constraint in favour of the sparsity constraint in LLC has the potential for a number of helpful properties. They may include: (a) better reconstruction: in contrast to a single basis codebook entry that VQ uses to represent each descriptor, LLC employs multiple bases. This means that while in the former approach similar descriptors may have very different codes, in the latter correlations between descriptors can be captured and the bases are shared; (b) locally smooth sparsity: reconstruction error is reduced in LLC through the use of multiple bases — *i.e.*, its explicit locality adaptor makes sure that patches with similarities have correspondingly similar codes; (c) analytical solution: spare coding problems can usually be solved only numerically, whereas LLC allows an analytical derivation.

Let M and K denote the numbers of codebook entries and nearest codewords, computational complexity is reduced to $\mathcal{O}(M+K^2)$ from $\mathcal{O}(M^2)$, where $K \ll M$. A codebook learning step is built into LLC via an 'online method' of learning [15]; B is the initial codebook that has been trained via k-means clustering.

B is then updated in increments as the training descriptors are iterated. For each of these increments, single or small-batch examples x_i are taken up and used to provide the required solution, resulting in the LLC codes associated with the current codebook B. This process takes the forms of a feature selector, since it retains only a set of bases B_i, the corresponding weights of which exceed a pre-set constant and refits x_i but omitting locality constraints. The code thereby generated can be employed to update the basis using a gradient descent. In brief, the LLC algorithm can be summed up in the following distinct steps:

1. Using 2D-SIFT descriptors, the local descriptor is identified within a dense image grid;
2. The local descriptor is translated into SC using LLC's nonlinear descriptor coding;
3. The representational outcome (feature representation) is generated by SC being submitted to multi-scale spatial pyramid max pooling.

There are clear benefits to this approach, *e.g.*, speed, simplicity and scalability, while providing the comparable performance to the SPM with SC [14].

3.3 Spatio-temporal LLC

The approach contains three steps; capturing video events with spatio-temporal local descriptors X, learning the locality-constrained sparse code S, and finally learning and optimising the codebook B (*c.f.*, Figure 1).

Spatio-temporal Interest Points. Interest points are detected using the ST-SIFT detector. To describe the region around the detected points, the 3D-HOG (histogram of oriented gradients) descriptor developed by Scovanner *et al.* [9] is used, calculating the spatio-temporal gradient for each pixel in the given cuboid. The approach leads to local regions that are invariant to scale and location in both the spatial and the temporal domains.

Learning Locality-Constrained Sparse Coding. The approach follows Wang *et al.* [15]; the criteria for the spatio-temporal LLC is

$$\min_{S} \sum_{i=1}^{N} \|x_i - Bs_i\|^2 + \lambda \|d_i \odot s_i\|^2 \quad st. \quad 1^{\top} s_i = 1, \, \forall i \tag{2}$$

where $\odot$ is the element-wise multiplication. The locality constrained parameter d_i represents every basis vector with different freedom based on its similarity to the spatio-temporal descriptor x_i:

$$d_i = \exp\left(\frac{dist(x_i, B)}{\sigma}\right) \quad st. \quad dist(x_i, B) = [dist(x_i, b_1), \ldots, dist(x_i, b_M)]^T \tag{3}$$

where $dist(x_i, b_1)$ is the Euclidean distance between the spatio-temporal descriptor and the basis codebook, and σ is the weight to control the locality parameter.

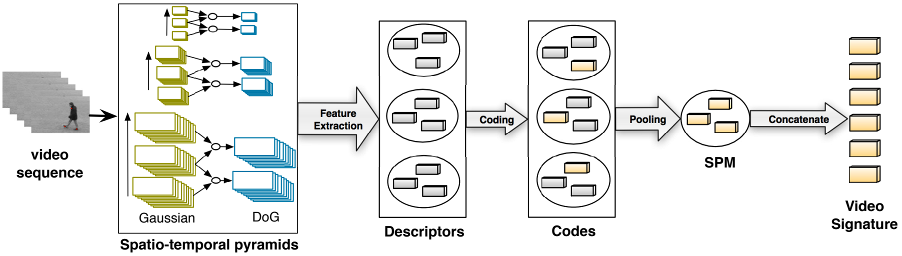

Fig. 1. ST-SIFT is combined with LLC

Codebook Optimisation. Given a set of n-dimensional spatio-temporal descriptors, $x_1, x_2, \ldots, x_m$, we generate an initial codebook using the k-means clustering method. Optimisation is performed so that the product of LLC coefficients and codebook basis should best approximate each spatio-temporal descriptor. The objective function can be defined as [15]:

$$\underset{S,B}{\operatorname{argmin}} \sum_{i=1}^{N} \{ \|x_i - Bs_i\|^2 + \lambda \|d_i \odot s_i\|^2 \quad st. \quad 1^\top s_i = 1, \forall i \quad \|b_j\|^2 \leq 1, \forall j \} \tag{4}$$

This is a convex problem in B only or S but not in both together, and can be iteratively solved by the coordinate descent method:

1. Initialise the dictionary B with the codebook generated by clustering:
$$B \leftarrow B_{init} \tag{5}$$
2. For each spatio-temporal descriptor x_i, compute the new LLC coefficient s_i using the current B:
$$s_i \leftarrow \underset{s}{\operatorname{argmax}} \|x_i - Bs\|^2 + \lambda \|d \odot s\|^2 \quad st. \quad 1^\top s = 1 \tag{6}$$
3. Update the current dictionary only if the computed LLC coefficient value is greater than a predefined threshold:
$$\triangle B_i \leftarrow -2\widetilde{s}_i(x_i - B_i\widetilde{s}_i), \quad \mu \leftarrow \sqrt{\frac{1}{i}}, \quad B_i \leftarrow B_i - \frac{\mu \triangle B_i}{|\widetilde{s}_i|_2} \tag{7}$$
4. Project the computed dictionary onto the output matrix:
$$B(:, id) \leftarrow proj(B_i) \tag{8}$$

4 Experiments

We evaluated the spatio-temporal LLC (ST-LLC) using human action classification tasks.

4.1 Implementation

To extract interest points from the spatio-temporal video cube, ST-LLC was built on 2D-LLC based image classifier [15]. Firstly, the spatio-temporal regions around the interest points were described by the 3D-HOG. Publicly available code by Scovanner *et al.* [9] was used. For each interest point the descriptor length was 640-dimensional and was determined by the number of bins to represent angles, θ and ϕ, in the sub-histograms. In the SPM step, the ST-LLC codes were computed for each spatio-temporal sub-region and pooled together using multi-scale max pooling to create the corresponding pooled representation. We used 4×4, 2×2 and 1×1 sub-regions. The pooled features were then concatenated and normalised using ℓ^2-normalisation. The next step was dictionary generation; a sample of the generated descriptors for interest points were clustered to a pre-specified number of visual words. We used Elkan's k-means clustering algorithm from the *VLFeat toolbox* [18], which was faster than the standard Lloyd's k-means. A support vector machine (SVM) classifier was used to learn a model from signatures for each action. We used a one-vs-all trained linear SVM.

4.2 Experimental Setup

Four publicly available human actions datasets were employed for benchmark.

* KTH dataset [19] — six human actions (walking, jogging, running, hand-waiving, boxing and hand-clapping) with each action performed by 25 persons in four different scenarios with monotone background;
* Weizmann data [20] — human actions performed by nine actors in ten action categories: walking, running, jumping, gallop sideways, bending, one-hand-waving, two-hands-waving, jumping in place, jumping jack, and skipping. Each clip contains single person performing an action in static background;
* UCF sports data [21] — more realistic but challenging data collected from broadcast sport videos. Nine actions are publicly available: diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging, walking;
* Hollywood dataset [4] — samples from 32 real-world movies with human actions. This is a challenging dataset due to its dynamic background and camera motions, categorised to one or more of eight human actions: answer phone, get out car, hand shake, hug person, kiss, sit down, sit up, and stand up. The dataset is divided to a testing set of 211 clips collected from 20 movies and two training sets: '*automatic*' created by script-based action annotation, and '*clean*' labelled manually. We train the classifier using the *clean* training set containing 219 clips.

When constructing a Gaussian pyramid, the number of scales was set to three for each of four levels in the KTH and Weizmann datasets, and three for each of three levels in the UCF sports and Hollywood datasets. A leave-one-out cross validation was used. The following parameters were used: codebook size of 1024 words (the key parameter for dictionary training), the number of neighbours $K = 5$, $\lambda = 500$ in Equation (4) and $\sigma = 100$ in Equation (3).

Table 1. Comparison of ST-LLC and the conventional 2D-LLC

detector	descriptor	coding	KTH	Weizmann	UCF	Hollywood
ST-SIFT	3D-HOG	LLC	100%	100%	88.9%	50.0%
2D-DoG	2D-HOG	LLC	49.7%	48.0%	70.8%	19.4%

4.3 Results

Table 1 shows that the ST-SIFT detector followed by the 3D-HOG descriptor and LLC coding outperformed the conventional 2D-LLC (a combination of the original 2D DoG detector, the 2D-HOG descriptor by Lowe [8], and the spatial LLC). The accuracies with the ST-LLC representation were 100% for KTH, 100% for Weizmann, 88.89% for UCF sports and 50% for Hollywood dataset. This indicates that ST-SIFT with locality-constrained coding is able to (1) capture interest points that have vital information in both the spatial and the temporal domains and to (2) represent events in real video sequences.

4.4 Comparison with the Recent State-of-the-Art

The achieved results of the ST-LLC with the human action classification benchmarks are roughly comparable to the current state-of-the-art. The KTH and the Weizmann data are technically 'solved' datasets, as the classification accuracy of 100% was reported by several groups recently. Sun *et al.* [22] reached 100% for both datasets, while Weinland *et al.* [23], Schindler and Gool [24] and Yeffet and Wolf [25] achieved 100% for the Weizmann data. Previously Yao *et al.* [26] performed 97.8% with the Weizmann using the Hough transform voting framework. Campos *et al.* [27] achieved an accuracy of 96.7% with the Weizmann dataset and 93.5% with the KTH dataset by applying bags-of-words and spatio-temporal shapes to represent human actions. For the KTH data, Chen and Hauptmann [11] reported 95.8% and Wu *et al.* [28] resulted in 95.7%. Gilbert *et al.* [29] achieved 94.5% using a mined dense spatio-temporal features.

For the UCF sports data, the ST-LLC (88.9%) outperform most of the recently reported results including 79.2% by Yeffet and Wolf [25] and 80.0% by Campos *et al.*Ẇu *et al.* [28] applied a method based on Lagrangian particle trajectories and boosted the accuracy to 89.7%. The ST-LLC is in line with the state-of-the-art results published recently by Sun *et al.* [22] (86.9%), Weinland *et al.* [23] (87.7%) and Yao *et al.* [26] (86.6%). Finally for the Hollywood dataset, to our knowledge, the current best result was produced by Gilbert *et al.* [29] (53.5%) using the hierarchical data mining approach. Other reported results include Chen and Hauptmann [11] (30.9%), Klaser *et al.* [5] (24.7%), Laptev *et al.* [4] (27.0%) and Yeffet and Wolf [25] (36.8%).

5 Conclusion

In this paper we presented a spatio-temporal coding technique based on ST-SIFT descriptor for human action recognition task. The method extended the

LLC approach by utilising the ST-SIFT descriptor to densely extract salient feature points from a 3D signal. This produced a group of distinctive feature points which were invariant to scale, rotation and translation as well as robust to temporal variation. The experimental results showed that LLC with the ST-SIFT outperformed (or at least achieved as good as) the most of state-of-the-art approaches on human action classification benchmarks, including KTH, Weizmann, UCF sports and Hollywood datasets.

Acknowledgements. The first author would like to thank Umm Al-Qura University, Makkah, Saudi Arabia for funding this work as part of her PhD scholarship program.

References

1. Laptev, I.: On space-time interest points. International Journal of Computer Vision, 107–123 (2005)
2. Shechtman, E., Irani, M.: Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? IEEE Trans. Pattern Anal. Mach. Intell., 2045–2056 (2007)
3. Liu, J., Shah, M.: Learning human actions via information maximization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008)
4. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 16:1–16:43 (2008)
5. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC, pp. 995–1004. British Machine Vision Association (2008)
6. Wong, S., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: IEEE International Conference on Computer Vision, ICCV, pp. 3455–3460 (2007)
7. Willems, G., Tuytelaars, T., Van Gool, L.: An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 91–110 (2004)
9. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the international conference on Multimedia, pp. 357–360. ACM (2007)
10. Cheung, W., Hamarneh, G.: N-sift: N-dimensional scale invariant feature transform for matching medical images. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 720–723 (2007)
11. Chen, M.Y., Hauptmann, A.: Mosift: Recognizing human actions in surveillance videos. Transform, 1–16 (2009)
12. Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: Computer Vision and Pattern Recognition, CVPR, pp. 625–632 (2011)
13. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: IEEE International Conference on Computer Vision, ICCV, pp. 1458–1465 (2005)

14. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1794–1801 (2009)
15. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367 (2010)
16. Allaire, S., Kim, J., Breen, S., Jaffray, D., Pekar, V.: Full orientation invariance and improved feature selectivity of 3D sift with application to medical image analysis. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR (2008)
17. Lopes, A., Oliveira, R., de Almeida, J., de A Araujo, A.: Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition. In: XXII Brazilian Symposium on Computer Graphics and Image Processing, pp. 315–321 (2009)
18. Vedaldi, A., Fulkerson, B.: Vlfeat: an open and portable library of computer vision algorithms. In: Proceedings of the International Conference on Multimedia, pp. 1469–1472. ACM (2010)
19. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the International Conference on Pattern Recognition, pp. 32–36 (2004)
20. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: The IEEE International Conference on Computer Vision, ICCV, pp. 1395–1402 (2005)
21. Rodriguez, M., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008)
22. Sun, C., Junejo, I.N., Foroosh, H.: Action recognition using rank-1 approximation of joint self-similarity volume. In: IEEE International Conference on Computer Vision, ICCV, pp. 1007–1012 (2011)
23. Weinland, D., Özuysal, M., Fua, P.: Making Action Recognition Robust to Occlusions and Viewpoint Changes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 635–648. Springer, Heidelberg (2010)
24. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008)
25. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: IEEE International Conference on Computer Vision, ICCV, pp. 492–497 (2009)
26. Yao, A., Gall, J., Van Gool, L.: A hough transform-based voting framework for action recognition. In: Computer Vision and Pattern Recognition, CVPR, pp. 2061–2068 (2010)
27. de Campos, T., Barnard, M., Mikolajczyk, K., Kittler, J., Yan, F., Christmas, W., Windridge, D.: An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In: IEEE Workshop on Applications of Computer Vision, WACV, pp. 344–351 (2011)
28. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: IEEE International Conference on Computer Vision, ICCV, pp. 1419–1426 (2011)
29. Gilbert, A., Illingworth, J., Bowden, R.: Fast realistic multi-action recognition using mined dense spatio-temporal features. In: IEEE International Conference on Computer Vision, ICCV, pp. 925–931 (2009)