Statistics of Pairwise Co-occurring Local Spatio-Temporal Features for Human Action Recognition

Piotr Bilinski and Francois Bremond

INRIA Sophia Antipolis, STARS team
2004 Route des Lucioles, 06902 Sophia Antipolis, France
{Piotr.Bilinski,Francois.Bremond}@inria.fr
www.inria.fr

Abstract. The bag-of-words approach with local spatio-temporal features have become a popular video representation for action recognition in videos. Together these techniques have demonstrated high recognition results for a number of action classes. Recent approaches have typically focused on capturing global statistics of features. However, existing methods ignore relations between features and thus may not be discriminative enough. Therefore, we propose a novel feature representation which captures statistics of pairwise co-occurring local spatio-temporal features. Our representation captures not only global distribution of features but also focuses on geometric and appearance (both visual and motion) relations among the features. Calculating a set of bag-of-words representations with different geometrical arrangement among the features, we keep an important association between appearance and geometric information. Using two benchmark datasets for human action recognition, we demonstrate that our representation enhances the discriminative power of features and improves action recognition performance.

1 Introduction

In recent years, recognition of human actions has became one of the most popular topic in computer vision domain. It has many potential applications, such as video surveillance, video indexing, retrieving and browsing, sport event analysis, human-computer interface and virtual reality. Although various methods have been proposed and much progress has been made, action recognition still remains a challenging problem. The main issues are: variations in visual and motion appearance of both people and actions, occlusions, noise, enormous amount of video data and changes in viewpoint, scale, rotation and illumination.

Over the last decade, there have been many studies on the recognition of human actions in video. Most of the state-of-the-art approaches can be divided into four categories depending on the type of features used. The first group of methods uses silhouette information [1–5]. The second category of techniques analyses object or motion trajectories [6–9]. However, both of these groups require precise algorithms, which is often very difficult to achieve due to such challenges as:

low discriminative appearance, illumination changes, camera movement, occlusions, noise and drifting problems. The third group of methods uses local spatio-temporal features [10–14]. Local spatio-temporal features have recently become a very popular video representation for action recognition. They have demonstrated promising recognition results for a number of action classes. They are able to capture both motion and visual appearance. Moreover, they are robust to scale variations, viewpoint changes and background clutter. Over the last decade, many algorithms have been proposed to detect local spatio-temporal interest points (e.g. Harris3D [10], Cuboid [15], Hessian [16] or Dense sampling [17]) and represent them using spatio-temporal descriptors (e.g. HOG [18], HOG3D [12], HOF [18], Cuboid [15] or ESURF [16]). One of the most frequently used detectors in the literature is the Harris3D [10], which is a space-time extension of the Harris operator. This algorithm is usually applied with Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow (HOF) descriptors [18]. The former describes local visual appearance and the latter characterizes local motion appearance of an interest point.

Local spatio-temporal features have been mostly used with a bag-of-words model. Together, these techniques have shown to achieve high recognition rate across various datasets. The bag-of-words model encodes global statistics of features, computing histogram of feature occurrences in a video sequence. This technique also has its own limitations. One of the main drawbacks of the bag-of-words model is that it ignores local pairwise relations among the features. To overcome this limitation, contextual features from the fourth category could be used. Contextual features can capture human-object interactions [19], scene context information [20, 21], figure-centric features [22–25], or pairwise relations between features [26, 24, 27]. Oikonomopoulos *et al.* [26] have proposed to construct class-specific codebooks of local features and encode spatial co-occurrences of pairs of codewords. Then, the action model is classified based on the probabilistic voting framework. Liu *et al.* [14] have proposed to explore the correlation of the compact video-word clusters using a modified correlogram. Banerjee *et al.* [24] have proposed to learn local neighbourhood relationships between local features, and train a CRF based human activity classifier. The neighbourhood relationships are modelled in terms of pairwise co-occurrence statistics. However, these methods are restricted using discriminative power of individual local features and ignoring association between appearance and geometric information. Thus, the performance of these techniques mainly depends on a single type of applied features. Ta *et al.* [27] have proposed to encode both appearance and spatio-temporal relations of local features. However, by calculating two independent codebooks (one codebook per feature type), this method ignores important association between appearance and geometric information.

To differ from those ideas, we propose a novel representation based on local spatio-temporal features and bag-of-words technique. Recent methods have typically focused on capturing global statistics of features. However, existing approaches ignore relations between the features, and thus may not be discriminative enough. Therefore, we propose a novel feature representation which

captures statistics of pairwise co-occurring local spatio-temporal features. Our representation captures not only global distribution of features but also focuses on geometric and appearance (both visual and motion) relations among the features. Calculating a set of bag-of-words representations with different geometrical arrangement among the features, we keep an important association between appearance and geometric information. We evaluate our approach on two publicly available datasets for human action recognition (KTH and UCF-ARG datasets). We show that the proposed representation enhances the discriminative power of local features and improves action recognition performance.

The rest of the paper is organized as follows. In section 2, we present our novel action recognition approach. In section 3, we present obtained results from our extensive set of experiments. Finally, in section 4, we conclude with future directions of work.

2 Proposed Approach

We propose a novel feature representation which captures statistics of pairwise co-occurring local spatio-temporal features. Firstly, we detect local interest points and capture both motion and visual appearance around extracted points. Then, we create a set of bag-of-words representations with different geometrical arrangement among the features. Our representation captures not only global distribution of features but also focuses on geometric and appearance (both visual and motion) relations among the features. Moreover, calculating a set of bag-of-words representations, we keep an important association between geometric and appearance information. The technique presented in this section enhances the discriminative abilities of features and improves action recognition performance.

2.1 Feature Extraction

Local spatio-temporal features have demonstrated high recognition results for a number of action classes. Therefore, we use them as basic features for our approach.

For each video sequence, we extract local spatio-temporal points of interest and their local spatio-temporal descriptors. To detect interest points, we use the sparse Harris3D corner detector [10]. To enhance the probability of capturing relevant information, we apply an algorithm searching over multiple spatial and temporal scales. Then, for each detected point, we compute HOG and HOF descriptors.

We highlight here that, all the mentioned algorithms in this section were selected based on their use in the literature and provide a good baseline for comparison with state-of-the-art techniques. However, our action representation method is independent of the type of detector and descriptor, and can be used together with any other algorithm.

2.2 Statistics of Pairwise Co-occurring Local Spatio-Temporal Features

In this section, we present our novel feature representation which captures statistics of pairwise co-occurring local spatio-temporal features. The following steps are applied for each video sequence independently.

Firstly, we extract local spatio-temporal interest points $\mathbb{P} = \{P_1, ..., P_n\}$ (where $P_l = (x_l, y_l, t_l)$) in a video sequence (Section 2.1). Then, for every point P_i we find its n-nearest neighbour points from the extracted set of points $\mathbb{P}$:

$$\mathbb{F}(\mathbb{P}) = \{(P_i, P_j) \in \mathbb{P}^2 : i_{nn}(\mathbb{P}, i, j) \leq n\}, \tag{1}$$

where $i_{nn}(\mathbb{P}, i, j) = m$ means that point P_j is the m-th nearest neighbour in order to point P_i. To calculate the distance between two points, we use the Euclidean metric.

Then, to differentiate pairs of points between those that are close to each other from those that are far away from each other, we split the set $\mathbb{F}(\mathbb{P})$ to several smaller subsets based on the value of the function i_{nn}:

$$\mathbb{S}(\mathbb{P}, a, b) = \{(P_i, P_j) \in \mathbb{F}(\mathbb{P}) : i_{nn}(\mathbb{P}, i, j) \in \langle a, b)\}, \tag{2}$$

where $0 \leq a \leq b \leq n$.

As we mentioned in the previous section, each point P_i is represented not only by its $3D$ position but also by the HOG-HOF descriptor. For simplicity, we indicate a HOG-HOF descriptor assigned to point P_i as $\mathfrak{D}(P_i)$. Therefore, to capture appearance (both visual and motion) relationship between two points, we represent each pair of points from the set $\mathbb{S}(\mathbb{P}, a, b)$ as a concatenation of their descriptors:

$$\mathbb{D}(\mathbb{P}, a, b) = \{\mathfrak{D}(P_i)||\mathfrak{D}(P_j) : (P_i, P_j) \in \mathbb{S}(\mathbb{P}, a, b)\}, \tag{3}$$

where $||$ is the concatenation operator.

Finally, we represent each video sequence as a collection of sets of features $\mathbb{D}(\mathbb{P}, a, b)$:

$$\mathbb{V}(\mathbb{P}, \mathbb{K}) = (\mathbb{D}(\mathbb{P}, k_1, k_2), \mathbb{D}(\mathbb{P}, k_2, k_3), ..., \mathbb{D}(\mathbb{P}, k_{|\mathbb{K}|-1}, k_{|\mathbb{K}|})), \tag{4}$$

where $\mathbb{K} = (k_1, k_2, ..., k_{|\mathbb{K}|})$. These sets vary in different geometrical arrangement among the features.

Our novel representation of features captures geometric and appearance (both visual and motion) relations among the features. Moreover, calculating sets of features $\mathbb{D}(\mathbb{P}, a, b)$, we keep an important association between geometric and appearance information. Thus, by using suitable designed features, we are able to overcome the limitation of the bag-of-words approach.

2.3 Action Representation

To represent videos, we apply the bag-of-words model for each feature class (*i.e.* HOG-HOF and $\mathbb{D}(\mathbb{P},.,.)$) independently. We construct visual vocabularies from training videos clustering computed features. Then, we assign each feature to its closest visual world. The obtained histograms of visual world occurrences over video forms the final representation.

The amount of features $\mathbb{D}(\mathbb{P},.,.)$ extracted from all the training videos can be large. Therefore, to speed-up the approach, we propose to perform clustering in the following hierarchical manner. In the first step, we process each video sequence independently. To reduce the computational cost, we limit the number of features for each video sequence to F_{MAX} using random sampling. Then, the obtained features are clustered. In the second step, we process all the training videos together and re-cluster all the obtained groups of features to create a final codebook representation.

2.4 Action Classification

To recognize an action, we use Multiple Kernel Learning (MKL) formulated for multi-class classification problem. We use MKL because it provides a natural method to combine different types of features. Given a list of base kernel functions, MKL searches for their linear combination which maximizes a performance measure. MKL considers a convex combination of n kernels:

$$K(H_i, H_j) = \sum_{z=1}^{n} \beta_z K_z(H_i, H_j), \tag{5}$$

with $\beta_z \geq 0$ and $\sum_{z=1}^{n} \beta_z = 1$.

To compare two m-bins histograms $H_i = [H_i(1), ..., H_i(m)]^T$ and $H_j = [H_j(1), ..., H_j(m)]^T$, we apply a χ^2 distance:

$$\chi^2(H_i, H_j) = \frac{1}{2} \sum_{z=1}^{m} \left(\frac{(H_i(z) - H_j(z))^2}{H_i(z) + H_j(z)} \right) \tag{6}$$

This distance is then converted into a χ^2 kernel using a multi-channel generalized Gaussian kernel:

$$K_z(H_i, H_j) = exp(-\frac{1}{A_z} \chi^2(H_i, H_j)) \tag{7}$$

where A_z is the normalization parameter set as in [18].

3 Experiments

Our experiments demonstrate the effectiveness of the proposed representation for a various of action categories. We evaluate our approach on two benchmark datasets for human action recognition - KTH and UCF-ARG datasets. Sample frames from video sequences of these datasets are presented in Figure 1. The performed experiments demonstrate that our representation enhances the discriminative power of features and improves action recognition accuracy.

Fig. 1. Sample frames from video sequences of the KTH (first row) and UCF-ARG (second row) datasets.

3.1 Implementation Details

In order to quantize local features, we use the k-means clustering technique. We use the L_2 norm metric to calculate the distance between features and visual words. We set the maximum amount of features extracted from a single video sequence to $F_{MAX} = 10^5$, which is a good compromise between the amount of data obtained from a video sequence and the time needed for clustering. We set the size of the codebook to 1000, which has shown empirically to give good results. In order to create statistics of pairwise co-occurring local spatio-temporal features, we set the parameter $\mathbb{K}$ to $(1, 2, 4, 8, 16)$, which has shown empirically to give good results.

In all our experiments, we apply the cross-validation technique to both gauge the generalizability of the proposed approach, and select the most discriminative statistics of pairwise co-occurring local spatio-temporal features. We use the Leave-One-Out Cross-Validation (LOOCV) technique, where videos of one person are used as the validation data, and the remaining videos as the training data. This is done repeatedly so that the videos of each person are used once as the validation data.

3.2 KTH Dataset

The KTH [28][1] dataset contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. Each action is performed several times by 25 different subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). The dataset contains 599 video files. All sequences were recorded with 25 fps frame rate.

[1] http://www.nada.kth.se/cvap/actions/

The dataset contains a set of challenges like: scale changes, illumination variations, shadows, different scenarios, cloth variations, inter and intra action class speed variations and low resolution (160×120 pixels spatial resolution).

We follow recent evaluations on the KTH dataset [29–33] using LOOCV scheme. In general, LOOCV assesses the performance of an approach with much more reliability than splitting-based evaluation schemes because it is much more comprehensive. Results from the experiments are presented in Table 1. Comparison of our approach with state-of-the-art methods in the literature using LOOCV technique is presented in Table 2. For scenarios $s1$, $s2$, $s3$ and $s4$, our approach obtains the recognition rate of 98.67%, 95.33%, 93.20% and 98.00% respectively. Overall, our approach obtains 96.30% recognition rate. The results clearly show that our representation enhances the discriminative power of features, improves action recognition performance and outperforms state-of-the-art techniques.

KTH	Recognition Rate
s1	98.67%
s2	95.33%
s3	93.20%
s4	98.00%
s1-s4	**96.30%**

Table 1. KTH dataset: Evaluation of our approach. The table shows the recognition rate overall and for each scenario independently.

Method	Year	Recognition Rate
Ta *et al.* [27]	2010	93.0%
Liu *et al.* [29]	2009	93.8%
Wu *et al.* [30]	2011	94.5%
Kim *et al.* [31]	2007	95.33%
Wu *et al.* [32]	2011	95.7%
Lin *et al.* [33]	2011	95.77%
Our method		**96.30%**

Table 2. KTH dataset: Comparison of our approach with state-of-the-art methods in the literature.

3.3 UCF-ARG Dataset

The UCF-ARG[2] (University of Central Florida - Aerial camera, Rooftop camera and Ground camera) dataset is a multiview human action dataset. It contains 12 actors performing ten types of human activities: boxing, carrying, clapping, digging, jogging, open-close trunk, running, throwing, walking and waving. Except for open-close trunk, all the other actions are performed 4 times by each actor in different directions. The open-close trunk action is performed 3 times. In total, we use 468 video sequences from the ground camera. The dataset is recorded using a high-definition camcorder (Sanyo Xacti FH1A camera) with 60 fps frame rate and spatial resolution of 1920×1080 pixels.

The dataset contains a set of challenges like: different shapes, sizes and ethnicities of people, scale changes, shadows, cloth variations, inter and intra action class speed variations, and different scenarios.

To the best of our knowledge there are not publicly available results for this dataset. Therefore, we compare our approach with popular baseline approach [18]. We use Harris3D to detect local spatio-temporal interest points and HOG-HOF descriptors to represent 3D video patches in the neighbourhood of detected points. Then, we apply bag-of-words model to represent video sequences and SVM for classification. The results from the experiments are presented in Table 3. The baseline approach obtains 80.98% recognition rate. Our proposed statistics of pairwise co-occurring local spatio-temporal features improve action recognition rate achieving 82.05% accuracy. We observe that also on this dataset, our representation enhances the discriminative power of local features and improves action recognition performance.

Method	Year	Recognition Rate
Laptev *et al.* [18]	2008	80.98%
Our method		**82.05%**

Table 3. UCF-ARG dataset: Comparison of our approach with baseline state-of-the-art method.

4 Conclusions and Future Work

We have proposed a novel feature representation which captures statistics of pairwise co-occurring local spatio-temporal features. Our representation captures not only global distribution of features but also focuses on geometric and appearance (both visual and motion) relations among the features. Calculating a set of bag-of-words representations with different geometrical arrangement among the features, we keep an important association between appearance and geometric

[2] http://vision.eecs.ucf.edu/data/UCF-ARG.html

information. The proposed approach has been evaluated on two public benchmark datasets for human action recognition. Obtained results have demonstrated that our technique enhances the discriminative power of features and improves action recognition performance. In the future work, we intend to evaluate our technique using different interest point detectors and descriptors. We also intend to examine different machine learning techniques to combine various types of features.

Acknowledgements. This work was supported by the Région Provence-Alpes-Côte d'Azur. However, the views and opinions expressed herein do not necessarily reflect those of the financing institution.

References

1. Davis, J.: Hierarchical motion history images for recognizing human motion. In: IEEE Workshop on Detection and Recognition of Events in Video. (2001)
2. Ahad, M., Tan, J., Kim, H., Ishikawa, S.: Motion history image: its variants and applications. Machine Vision and Applications (2010)
3. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. CVIU (1999)
4. Kim, T.S., Uddin, Z. In: Silhouette-based Human Activity Recognition Using Independent Component Analysis, Linear Discriminant Analysis and Hidden Markov Model. InTech (2010)
5. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: ICCV. (2009)
6. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: ICCV. (2009)
7. Raptis, M., Soatto, S.: Tracklet descriptors for action modeling and video analysis. In: ECCV. (2010)
8. Kaaniche, M.B., Bremond, F.: Gesture recognition by learning local motion signatures. In: CVPR. (2010)
9. Wang, H., Klaser, A., Schmid, C., Cheng-Lin, L.: Action recognition by dense trajectories. In: CVPR. (2011)
10. Laptev, I.: On space-time interest points. IJCV (2005)
11. Rapantzikos, K., Avrithis, Y., Kollias, S.: Dense saliency-based spatiotemporal feature points for action recognition. In: CVPR. (2009)
12. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008)
13. Gilbert, A., Illingworth, J., Bowden, R.: Fast realistic multi-action recognition using mined dense spatio-temporal features. In: ICCV. (2009)
14. Liu, J., Shah, M.: Learning human actions via information maximization. In: CVPR. (2008)
15. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, in conjunction with ICCV. (2005)
16. Willems, G., Tuytelaars, T., Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV. (2008)

17. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009)
18. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
19. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: CVPR. (2007)
20. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: ICCV. (2007)
21. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)
22. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: CVPR. (2009)
23. Wang, J., Chen, Z., Wu, Y.: Action recognition with multiscale spatio-temporal contexts. In: CVPR. (2011)
24. Banerjee, P., Nevatia, R.: Learning neighborhood co-occurrence statistics of sparse features for human activity recognition. In: AVSS. (2011)
25. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR. (2010)
26. Oikonomopoulos, A., Patras, I., Pantic, M.: An implicit spatiotemporal shape model for human activity localisation and recognition. In: Workshop on Human Communicative Behaviour Analysis, in conjunction with CVPR. (2009)
27. Ta, A.P., Wolf, C., Lavoue, G., Baskurt, A., Jolion, J.M.: Pairwise features for human action recognition. In: ICPR. (2010)
28. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR. (2004)
29. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: CVPR. (2009)
30. Wu, X., Xu, D., Duan, L., Luo, J.: Action recognition using context and appearance distribution features. In: CVPR. (2011)
31. Kim, T.K., Wong, S.F., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: CVPR. (2007)
32. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: ICCV. (2011)
33. Jiang, Z., Lin, Z., Davis, L.: Recognizing human actions by learning and matching shape-motion prototype trees. PAMI (2011)