LETTER

Contextual Max Pooling for Human Action Recognition

Zhong ZHANG[†], *Member*, Shuang LIU[†a)], *and* Xing MEI[††,†††], *Nonmembers*

SUMMARY The bag-of-words model (BOW) has been extensively adopted by recent human action recognition methods. The pooling operation, which aggregates local descriptor encodings into a single representation, is a key determiner of the performance of the BOW-based methods. However, the spatio-temporal relationship among interest points has rarely been considered in the pooling step, which results in the imprecise representation of human actions. In this paper, we propose a novel pooling strategy named contextual max pooling (CMP) to overcome this limitation. We add a constraint term into the objective function under the framework of max pooling, which forces the weights of interest points to be consistent with their probabilities. In this way, CMP explicitly considers the spatio-temporal contextual relationships among interest points and inherits the positive properties of max pooling. Our method is verified on three challenging datasets (KTH, UCF Sports and UCF Films datasets), and the results demonstrate that our method achieves better results than the state-of-the-art methods in human action recognition.

key words: *contextual max pooling, human action recognition, spatio-temporal relationship*

1. Introduction

Recognizing human action has raised a great interest in computer vision and pattern recognition fields due to the requirements of real-world applications, such as video surveillance, human-computer interaction and video indexing. Recently, a lot of strategies have been proposed by researchers, such as 2-D shape matching [1], optical flow patterns [2], trajectory-based representation [3], spatio-temporal interest points [16] and attribute representation [4]. In particular, methods based on the spatio-temporal interest points with BOW model [5], [16] have shown promising performance. Since these approaches do not rely on preprocessing techniques, e.g. background modeling or body-part tracking, they are relatively robust to noise, background changes and illumination variation. The BOW-based methods follow a common work flow: they first extract local descriptors, and then encode these descriptors over some learned codebook or dictionary (coding step). Finally, the encodings are aggregated into a vector to represent the action video (pooling step). In this paper, we focus on the pooling step.

The common pooling methods used in human action recognition and image classification are sum pooling (or average pooling) and max pooling. Sum pooling lacks discrimination because it is strongly influenced by frequent descriptors whether they are beneficial to classification or not. The max pooling is only suitable for the coding strategies that rely on count statistics. Recently, some extensions to the sum pooling and max pooling have been proposed, one purpose of which is to produce better representations without losing too much information. To reduce the information loss, only close local descriptors should be pooled together in the geometric and descriptor domain [6]. Lin *et al.* [7] considered the latent image structure to learn the important pooling spatial regions for scene classification. Murray *et al.* [8] proposed a generalized version of max pooling which is applicable to variant coding strategies. However, none of the above pooling operations explicitly provide spatio-temporal relationships among interest points, which reflects both spatial relative layout of human body parts and temporal evolution of human poses.

In this paper, we propose a novel pooling strategy called contextual max pooling (CMP), which not only preserves the spatio-temporal relationships among interest points, but also inherits the positive properties of max pooling. A constraint, which assumes the weights of interest points to be consistent with their probabilities, is added into the objective function under the framework of max pooling. The probability of the local appearance descriptor in its location for an interest point is estimated by kernel density estimation (KDE) in the contextual domain.

The rest of this paper is organized as follows. Section 2 introduces the traditional max pooling and its extension. Section 3 presents our approach in detail. Section 4 shows the experimental results, demonstrating that our method outperforms the state-of-the-art methods on the KTH, UCF Sports and UCF Films datasets. Finally, in Sect. 5, we conclude the paper.

2. Traditional Max Pooling

Let $X = \{x_1, x_2, \ldots, x_N\}$ denote N local appearance descriptors of spatio-temporal interest points extracted from an action video, where $x_i \in \mathbb{R}^{D\times 1}$ is the local appearance feature for the i-th spatio-temporal interest point. Let $V = \{v_1, v_2, \ldots, v_N\}$ be a set of encodings of X, where $v_i \in \mathbb{R}^{M\times 1}$ indicates the encoding of x_i. After pooling operation, the en-

Manuscript received October 31, 2014.
Manuscript revised December 5, 2014.
Manuscript publicized January 19, 2015.

[†] The authors are with College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin, China.

[††] The author is with the Department of Computer Science, University at Albany, SUNY, New York, USA.

[†††] The author is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

a) E-mail: shuangliu.tjnu@gmail.com

DOI: 10.1587/transinf.2014EDL8221

coding set V is aggregated into a single vector h to represent an action video. One property of the traditional max pooling is that the inner product between the max pooling representation h^{max} and an interest point encoding v_i is a constant value [8], i.e. $v_i^T h^{max} = 1$, when max pooling operation is embedded in the BOW model (hard voting). Here, v_i is a binary vector with a single non-zero entry, and h^{max} is a binary vector to represent an action video where a 1 is indicative of the presence of the codeword. Murray and Perronnin [8] extended max pooling to a generalized form which is directly applicable to various coding strategies

$$V^T h = \mathbb{1}_N \tag{1}$$

where $\mathbb{1}_N \in \mathbb{R}^{N\times 1}$ is a vector of all ones. For computational convenient, we turn Eq. (1) to a least square problem. Additionally, it is beneficial to add a regularization term to obtain a stable solution. Hence, Eq. (1) is reformulated as

$$h^* = arg\min_h \parallel V^T h - \mathbb{1}_N \parallel^2 + \alpha \parallel h \parallel^2 \tag{2}$$

where α is a balancing parameter. Introducing $h = Vw$ into Eq. (2) according to the representer theorem [9], it can be rewritten as

$$w^* = arg\min_w \parallel V^T Vw - \mathbb{1}_N \parallel^2 + \alpha \parallel Vw \parallel^2 \tag{3}$$

where $w \in \mathbb{R}^{N\times 1}$ is the weight vector, and $w = [w_1, w_2, \ldots, w_N]^T$ in which w_i is the weight of i-th spatio-temporal interest point.

3. Approach

In this section, we first present the proposed contextual max pooling. Then, we introduce how to compute the probability of a spatio-temporal interest point.

3.1 Contextual Max Pooling

The traditional max pooling and its extension, however, neglect spatial and temporal relationships among interest points, which reflects both spatial relative layout of human body parts and temporal evolution of human poses. In this work, we explicitly consider spatio-temporal contextual information in the pooling step under the framework of max pooling, which is called contextual max pooling.

We use the joint probability density $p(x_i, s_i)$ to capture the spatial and temporal relationship among the interest points for each action video. Here $x_i \in \mathbb{R}^{D\times 1}$ is the i-th local appearance feature and s_i is the i-th interest point location, i.e. $s_i = (a_i, b_i, t_i)$ where a_i, b_i and t_i are horizontal, vertical, and temporal coordinates respectively. $p(x_i, s_i)$ indicates the probability of x_i occurring at location s_i and the computational details will be shown in the next subsection. If the $p(x_i, s_i)$ is larger, then this interest point has a greater influence on the final representation. In other words, we prefer that the wight of interest point w_i is consistent with the $p(x_i, s_i)$. With the above consideration, the constraint on the weights can be expressed as

$$\parallel w - q \parallel^2 = \sum_{i=1}^{N} (w_i - q_i)^2 \tag{4}$$

where $q = [q_1, q_2, \ldots, q_N]^T$ is the N-dimensional vector and $q_i = p(x_i, s_i)$. Equation (4) is actually served as a penalty term (as shown below), which results in a high penalty if the weights of interest points are different from the probabilities at their location. In other words, it forces the wights to be consistent with their probabilities.

We add the constraint in Eq. (4) into Eq. (3) to obtain our CMP. The optimization problem can be expressed as

$$w^* = arg\min_w \parallel V^T Vw - \mathbb{1}_N \parallel^2 + \alpha \parallel Vw \parallel^2 + \beta \parallel w - q \parallel^2 \tag{5}$$

where $\beta \geq 0$ is a regularization parameter that controls the constraint on the weights of interest points. By adding the constraint on the weights, Eq. (5) not only inherits the positive properties of max pooling, but also explicitly considers the spatio-temporal relationship among interest points. It should be noticed that when $\beta = 0$, our model will degenerate into the generalized max pooling [8]. Furthermore, when $\beta = 0$ and the pooling operation is embedded in the BOW model (hard voting), our model will degenerate into the traditional max pooling [10]. Our CMP model has a closed-form solution

$$w = (K + \alpha E + \beta K^{-1})^{-1} (\beta K^{-1} q + \mathbb{1}_N) \tag{6}$$

where $K = V^T V$ is the $N \times N$ similarity kernel matrix and E is the identity matrix of size $N \times N$ with ones on the main diagonal and zeros elsewhere. From Eq. (6), we can see that the weights of interest points for each action video are determined by their appearance and spatio-temporal contextual relationship which are considered by K and q respectively.

3.2 Computing the Probability $p(x_i, s_i)$

Given a spatio-temporal point (x_i, s_i), its surrounding spatio-temporal area is called contextual domain [11] which is a cube with a predefined side length shown as the cube in Fig. 1. Let $B = [b_1, b_2, \ldots, b_M]$ denote the codebook with M clustering centers. The context for the codeword b_i is defined as

$$U_i = \{(x', s') | x' \in b_i, s' \in \Omega(s_i)\} \tag{7}$$

where $\Omega(s_i)$ is the contextual domain of interest point (x_i, s_i). Equation (7) indicates that a spatio-temporal point (x', s'), whose local appearance feature is x' and location is s', belongs to U_i when x' is the nearest to the codeword b_i and s' is in the contextual domain.

We calculate $p(x_i, s_i)$ in its contextual domain:

$$p(x_i, s_i) \doteq p(b_i, s_i) = p(s_i | b_i) \cdot p(b_i) \tag{8}$$

where x_i is the nearest to the codeword b_i. The prior prob-

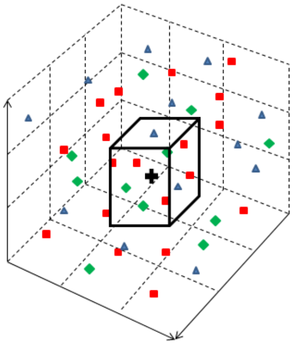


Fig. 1 The black cube is the contextual domain of the black cross interest point.

ability density $p(b_i)$ can be estimated in the contextual domain U_i, i.e., the percentage of codeword b_i in the contextual domain, and the $p(s_i|b_i)$ can be computed by the kernel density estimation:

$$p(s_i|b_i, U_i) \propto \sum_{s_k \in U_i} \Psi(s_i, s_k) \tag{9}$$

where $\Psi(s_i, s_k)$ is a 3D Gaussian kernel:

$$\Psi(s_i, s_k) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(s_i - s_k)^T \Sigma^{-1} (s_i - s_k)\right\} \tag{10}$$

where s_k is the point location in U_i and Σ is the covariance matrix. We assume that horizontal, vertical and temporal coordinates are independent. Thus

$$\Sigma = diag(\sigma_u^2, \sigma_w^2, \sigma_t^2) \tag{11}$$

where σ_u, σ_w, and σ_t are spatial and temporal scale parameters.

4. Experimental Results

We verify our CMP method on several benchmark datasets: KTH dataset [12], UCF Sports dataset [13], and UCF Films dataset [13]. We compare our algorithm with relevant baselines and other excellent algorithms on human action recognition. There are three algorithms as our relevant baselines: sum pooling (SP), max pooling (MP) and generalized max pooling (GMP) [8]. In our experiments, we adopt the Harris 3-D corner [14] to detect interest points from action videos. For each interest point, the histogram of oriented gradients (HOG) and histogram of optical flow (HOF) are used as local appearance descriptors. Then, we encode the local appearance descriptors by utilizing LLC [15]. Finally, we use pooling operation to aggregate the encodings and represent an actio video as a feature vector. We set the side length of contextual domain to 35 and the number of codebook to 4000 using k-means clustering algorithm.

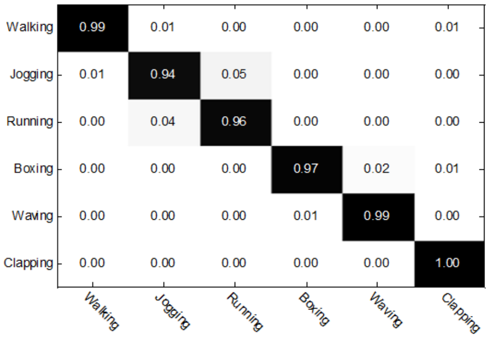
The KTH dataset is a widely used action dataset which contains six human action categories. They are performed by 25 subjects under four different scenarios, resulting in a total of 599 video clips. We adopt the leave-one-out cross validation strategy, specifically 24 videos of actors as training and the rest one as test videos. The choices of α and β in Eq. (5) have an impact on the final performance. Table 1 shows the average accuracy values under different α and β, from which we can see that when $\alpha = 20$ and $\beta = 30$ results are the best. The paper mainly reports the results on the KTH dataset, and our experiments have shown that the conclusions can be generalized to the UCF Sports dataset and the UCF Films dataset as well.

Table 1 The action recognition accuracy values (%) under different α and β in the KTH dataset.

β \ α	10	20	30	40
20	94.6	94.8	94.2	93.6
30	96.2	**97.6**	96.8	95.3
40	95.1	94.4	93.6	91.9
50	94.5	94.7	93.2	92.1

Table 2 The comparison of our method with the state-of-the-art methods and the baseline methods on the KTH and UCF Sports datasets.

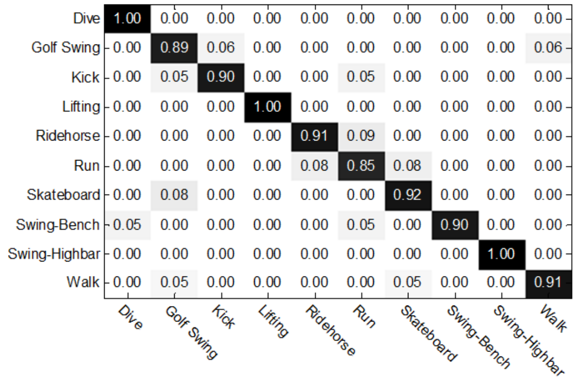
	KTH (%)	UCF (%)
Kovashka *et al.* [16]	94.5	87.3
Le *et al.* [17]	93.9	86.5
Wang *et al.* [18]	94.2	88.2
Jiang *et al.* [19]	95.8	88.0
Wang *et al.* [20]	93.3	-
Wu *et al.* [21]	97.0	90.7
SP	91.2	85.3
MP	92.4	87.3
GMP	93.8	88.0
CMP	**97.6**	**92.0**



	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	0.99	0.01	0.00	0.00	0.00	0.01
Jogging	0.01	0.94	0.05	0.00	0.00	0.00
Running	0.00	0.04	0.96	0.00	0.00	0.00
Boxing	0.00	0.00	0.00	0.97	0.02	0.01
Waving	0.00	0.00	0.00	0.01	0.99	0.00
Clapping	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 2 Confusion table of our method on the KTH database.

The average accuracy values on the KTH dataset are shown in Table 2, and the confusion table of recognition results on the KTH dataset is shown in Fig. 2. With the optimal parameters, our CMP method achieves the best accuracy value of 97.6% on the KTH dataset. Furthermore, the following four points can be drawn through analyzing the experimental results. First, comparing CMP approach with GMP approach, we can see that the former is 3.8% higher than the latter one on the accuracy. It shows that the spatio-temporal relationship explicitly considered by our



	Dive	Golf Swing	Kick	Lifting	Ridehorse	Run	Skateboard	Swing-Bench	Swing-Highbar	Walk
Dive	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Golf Swing	0.00	0.89	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.06
Kick	0.00	0.05	0.90	0.00	0.00	0.05	0.00	0.00	0.00	0.00
Lifting	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Ridehorse	0.00	0.00	0.00	0.00	0.91	0.09	0.00	0.00	0.00	0.00
Run	0.00	0.00	0.00	0.00	0.08	0.85	0.08	0.00	0.00	0.00
Skateboard	0.00	0.08	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00
Swing-Bench	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.90	0.00	0.00
Swing-Highbar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Walk	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.91

Fig. 3 Confusion table of our method on the UCF Sports database.

Table 3 The comparison of our method with the state-of-the-art methods and the baseline methods on the UCF Films datasets.

	Kiss (%)	Slap (%)	Average (%)
Rodriguez *et al.* [13]	66.4	67.2	66.8
Yeffet *et al.* [22]	77.3	84.2	80.7
Wu *et al.* [21]	97.6	94.4	96.0
CMP	97.8	94.6	**96.2**

CMP approach can improve the performance. Second, the CMP, GMP and MP methods outperform the SP method. It demonstrates that max pooling and its extensions can capture the discriminative information when aggregating the encodings. Third, our CMP gains 6.4% accuracy rate over SP. It is because our CMP not only inherits the positive properties of max pooling but also preserves spatio-temporal relationships among the interest points. Finally, from the confusion table, we can see that leg-related actions (Running and Jogging) are prone to be misclassified. We think the possible reason may be that they always exhibit similar context and appearance.

The UCF Sports dataset contains 150 sports videos of ten action categories. This dataset represents a natural pooling of actions featured in a wide range of scenes and viewpoints, so the videos exhibit great intra-class variation. We take the leave-one-out cross validation, namely cycling each sample as a test video one at a time. The performances of different methods are shown in Table 2 and the confusion table of recognition results on the UCF Sports dataset is shown in Fig. 3. We can see that our CMP method outperforms the other three baselines and other state-of-the-art methods, reaching 91.9% on the UCF Sports dataset. We draw the similar conclusions with that on the KTH datasets, which proves the effectiveness of our CMP on the realistic and complicated action dataset.

The UCF Films dataset provides a representative pool of natural samples of action classes, including kissing and slapping. There are 92 videos of kissing and 112 videos of slapping. These actions are performed in classic old movies. We adopt leave-one-out cross validation and the results are shown in Table 3. Once again, we prove the effectiveness of our algorithm on this dataset.

5. Conclusions

In this paper, we propose a novel pooling strategy called CMP to overcome the limitation of traditional max pooling. The CMP explicitly considers the spatio-temporal contextual relationships among interest points, which can provide more accurate pooling than the traditional max pooling. The proposed method has been validated on three challenging datasets, and the experimental results clearly demonstrate the superiority of our method over previous methods in human action recognition.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant No. 61401309, and No. 61271430 and Doctoral Fund of Tianjin Normal University under Grant No. 5RL134.

References

[1] Z. Lin, Z. Jiang, and L.S. Davis, “Recognizing actions by shape-motion prototype trees,” ICCV, pp.444–451, 2009.

[2] A. Efros, A. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” ICCV, pp.726–733, 2003.

[3] M. Raptis and S. Soatto, “Tracklet descriptors for action modeling and video analysis,” ECCV, pp.577–590, 2010.

[4] J. Liu, B. Kuipers, and S. Savarese, “Recognizing human actions by attributes,” CVPR, pp.3337–3344, 2011.

[5] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, “Action recognition using context-constrained linear coding,” IEEE Signal Process. Lett., vol.19, no.7, pp.439–442, 2012.

[6] Y.-L. Boureau, N.L. Roux, F. Bach, J. Ponce, and Y. LeCun, “Ask the locals: Multi-way local pooling for image recognition,” ICCV, pp.2651–2658, 2011.

[7] D. Lin, C. Lu, R. Liao, and J. Jia, “Learning important spatial pooling regions for scene classification,” CVPR, pp.3726–3733, 2014.

[8] N. Murray and F. Perronnin, “Generalized max pooling,” CVPR, pp.2473–2480, 2014.

[9] B. Schölkopf, R. Herbrich, and A.J. Smola, “A generalized representer theorem,” Computational Learning Theory, pp.416–426, 2001.

[10] Y.-L. Boureau, J. Ponce, and Y. LeCun, “A theoretical analysis of feature pooling in visual recognition,” ICML, pp.111–118, 2010.

[11] Y. Wu and J. Fan, “Contextual flow,” CVPR, pp.33–40, 2009.

[12] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” ICPR, pp.32–36, 2004.

[13] M.D. Rodriguez, J. Ahmed, and M. Shah, “Actio mach: A spatio-temporal maximum average correlation height filter for action recognition,” CVPR, pp.1–8, 2008.

[14] I. Laptev, “On space-time interest points,” Int. J. Comput. Vis., vol.64, no.2-3, pp.107–123, 2005.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” CVPR, pp.3360–3367, 2010.

[16] A. Kovashka and K. Grauman, “Learning a hierarchy of discriminative space-time neighborhood features for human action recognition,” CVPR, pp.2046–2053, 2010.

[17] Q. Le, W. Zou, S. Yeung, and A. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” CVPR, pp.3361–3368, 2011.

[18] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, “Action recognition by dense trajectories,” CVPR, pp.3169–3176, 2011.

[19] Z. Jiang, Z. Lin, and L.S. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.3, pp.533–547, 2012.

[20] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3D parts for human motion recognition," CVPR, pp.2674–2681, 2013.

[21] B. Wu, C. Yuan, and W. Hu, "Human action recognition based on context-dependent graph kernels," CVPR, pp.2609–2616, 2014.

[22] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," ICCV, pp.492–497, 2009.