Human Action Recognition Based on Oriented Motion Salient Regions

Baoxin Wu[1], Shuang Yang[1], Chunfeng Yuan[1], Weiming Hu[1], and Fangshi Wang[2]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{bxwu,syang,cfyuan,wmhu}@nlpr.ia.ac.cn
[2]School of Software, Beijing Jiaotong University, Beijing, China
{fshwang}@bjtu.edu.cn

Abstract. Motion is the most informative cue for human action recognition. Regions with high motion saliency indicate where actions occur and contain visual information that is most relevant to actions. In this paper, we propose a novel approach for human action recognition based on oriented motion salient regions (OMSRs). Firstly, we apply a bank of 3D Gabor filters and an opponent inhibition operator to detect OMSRs of videos, each of which corresponds to a specific motion direction. Then, a new low-level feature, named as oriented motion salient descriptor (OMSD), is proposed to describe the obtained OMSRs through the statistics of the texture in the regions. Next, we utilize the obtained OMSDs to explore the oriented characteristics of action classes and generate a set of class-specific oriented attributes (CSOAs) for each class. These CSOAs provide a compact and discriminative middle-level representation for human actions. Finally, an SVM classifier is utilized for human action classification and a new compatibility function is devised for measuring how well a given action matches to the CSOAs of a certain class. We test the proposed approach on four public datasets and the experimental results validate the effectiveness of our approach.

1 Introduction

Traditional approaches for human action recognition are based on either local or global features. The former [5, 2, 3] is usually extracted from a sparse set of local salient regions and tends to lose useful global information about the action. Contrastively, the latter [6, 7] treats the video as a whole. It contains all the information of the sequence but is sensitive to occlusion and background variation. Regions with high motion saliency are of great significance because they indicate where actions occur in videos and contain the most relevant information about the actions. In this paper, we propose a novel approach for human action recognition based on the oriented motion salient regions (OMSRs).

Much effort has been devoted to estimate motion using successive frames, but detecting motion in specific directions is still a challenge. In this paper, we apply a bank of 3D Gabor filters with multiple directions and an opponent inhibition

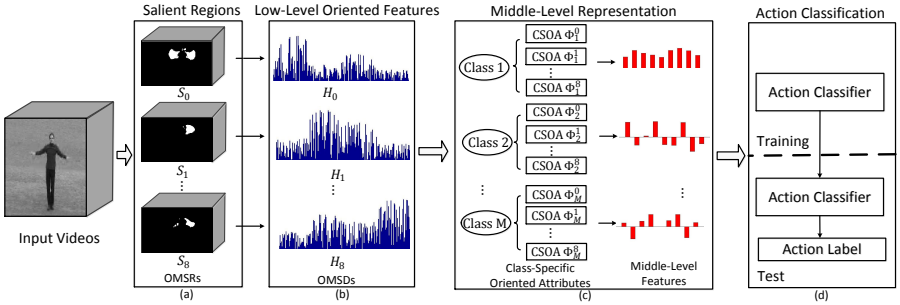

Fig. 1. The flowchart of our proposed approach. (a) The detected OMSRs. (b) The OMSDs extracted from OMSRs. (c) Learned CSOAs and the middle-level attriubte vectors. (d) Action classification based on CSOAs.

operator to detect the motion salient region in a video sequence. Specifically, the detected region is decomposed into a set of OMSRs, each of which corresponds to a specific motion direction.

We extract a new low-level descriptor from each OMSR, which is named as oriented motion salient descriptor (OMSD). The OMSDs are obtained by the statistics of the texture information in OMSRs and when computing them, a two-layer polar coordinate system is used to encode the spatial distribution information of the OMSRs. Each OMSD is actually a semi-holistic feature for human action representation because it contains all the information about an action in a specific direction.

Taking advantage of OMSDs, a series of class-specific oriented attributes (C-SOAs) are learnt for each action class. The CSOAs reflect the characteristics of action classes in different motion directions. Mapping an action represented by a set of low-level OMSDs into the space of CSOAs related to an class generates a middle-level attribute feature, which has high discriminative power for human action recognition. Finally, an SVM formulation is utilized for the action classification problem, where action classes are characterized by a set of class-specific attributes.

The flowchart of our proposed approach is shown in Figure 1. The main contributions of this paper are as follows:

- 3D Gabor filters, incorporating an opponent inhibition operator, are used to detect the oriented motion salient regions which contain the most relevant information about actions with respect to different motion directions.
- A new low-level descriptor OMSD, encoding both appearance and motion information of OMSRs, is proposed to represent the extracted regions.
- A set of CSOAs are learned for recognizing human actions. The CSOAs characterize action classes in different directions and provide an middle-level representation for human actions.

2 Related work

Features used for human action representation can be roughly categorized into two groups: local features and holistic features. The local feature based approaches represent an action as a sparse set of extracted local spatiotemporal features. Dollár *et al.* [2] experiment with both pixel gradients and optical flows to describe the extracted spacetime cuboids. Laptev *et al.* [3] apply histograms of gradients (HoG) and histograms of optical flows (HoF) as local descriptors, capturing both motion and structure information of local regions of interest points. These features are extracted from the local regions around the interest points and the information in other regions is usually ignored. Compared with these local features, the proposed OMSDs focus on the whole motion salient regions with respect to a specific direction, capturing more visual information which is useful for human action recognition.

The global feature based approaches represent an action by treating the video sequence as a whole. Bobick and Davids [6] introduce motion energy image (MEI) and motion history image (MHI) to describe the spatial distribution of motion energy in a given sequence. Wang *et al.* [7] construct average motion energy (AME) and mean motion shape (MMS) based on the human body silhouettes and shapes respectively, to characterize an action. Ikizler *et al.* [8] describe a pose as a histogram of oriented rectangles and then sum up the histograms from all frames to form a compact representation for the whole video sequence. These holistic features capture sufficient visual information. However, they are highly sensitive to shift and background variations. Our OMSDs, based on the oriented motion salient regions, are more robust to these variations.

In addition, inspired by the formulation on object recognition [11, 13], a number of researchers show great interest in the attribute based representation for human actions. Yao *et al.* [14] use attributes and parts for human action recognition in still images. Liu *et al.* [12] combine manually predefined attributes with data-driven ones obtained by clustering local features, and use a latent SVM to learn the importance of each part. There are two main differences between our CSOAs and these sematic attributes. First, our CSOAs are all learnt automatically without any manual annotation. Second, our CSOAs are class-specific attributes and they are more discriminative for classifying actions from different action classes.

Over the years, another group of methods perform human action recognition by aggregating the responses of 3D directional filters and these methods is quite related to our approach. Derpanis *et al.* [17, 22] propose to use 3D Gaussian third derivative filters for human action recognition. In their work a marginalization process is used to discount spatial orientation component. In our work, a more simple 3D Gabor filter and a opponent inhibition operator are designed for OMSR detection. On the other hand, we do not aggregate the motion energies of different directions simply, but give a mid-level representation of actions through exploiting the obtained CSOAs.

3 Oriented Motion Salient Regions

In this section, we introduce how to detect OMSRs in videos. Adelson and Bergen [10] have demonstrated that motion can be perceived as orientation in space-time and spatiotemporally oriented filters can be used to detect it. Inspired by this, we apply a bank of 3D Gabor filters [9] with multiple directions and an opponent inhibition operator for motion analysis in videos.

A 3D Gabor filter is formed as a product of a Gaussian window and complex sinusoid. It consists of two parts: the real part and the imaginary part. These two parts are defined as

$$g_r^{3d}(x,y,t) = \hat{g}(x,y,t)\cos[\frac{2\pi}{\lambda}(\eta_x x + \eta_y y + \eta_t t)], \tag{1}$$

and

$$g_i^{3d}(x,y,t) = \hat{g}(x,y,t)\sin[\frac{2\pi}{\lambda}(\eta_x x + \eta_y y + \eta_t t)], \tag{2}$$

respectively, where

$$\hat{g}(x,y,t) = \exp[-(\frac{x^2+y^2+t^2}{2\sigma^2})], \tag{3}$$

σ controls the scale of the Gaussian, λ is the wavelength of the sinusoidal factor, and (η_x, η_y, η_t), which satisfies $\eta_x^2 + \eta_y^2 + \eta_t^2 = 1$, determines the direction of the filter. The response of a 3D Gabor filter on a video sequence I is expressed as

$$R = (I * g_r^{3d})^2 + (I * g_i^{3d})^2, \tag{4}$$

where $*$ denotes the convolution operator. Through squaring and summing the outputs of two part filters which are 90 degrees out of phase, the 3D Gabor filter gives a phase-independent measurement of local motion strength.

To capture motions towards multiple directions, we design a filter bank which contains nine 3D Gabor filters with different directions. Let $\{g_k^{3d}\}_{k=0}^{8}$ denote the filters in the bank. These filters are sensitive to motions with any of directions: flicker, up, down, left, right and four diagonals. Table 1 shows the filters and their corresponding motions. By convoluting a video sequence with these 3D Gabor filters, we obtain the responses $\{R_k\}_{k=0}^{8}$, which are actually a series of oriented motion energy measurements on this sequence.

Each 3D Gabor filter responds to motion with a specific direction independently. However, the motion detection should be inherently opponent [10]. That is motions with two opposite directions cannot occur at the same place and time within the same frequency band. Accordingly, we apply an opponent inhibition operator on the original responses $\{R_k\}_{k=0}^{8}$, which will decrease the influence of opposite motion. The opponent inhibition operator is defined as the half-wave-rectified difference between the oriented motion energies corresponding to opposite motion directions

$$\begin{aligned} &\bar{R}_0(\boldsymbol{x}) = R_0(\boldsymbol{x}), \\ &\bar{R}_i(\boldsymbol{x}) = |R_i(\boldsymbol{x}) - aR_{i+4}(\boldsymbol{x})|^+, \quad 1 \le i \le 4, \\ &\bar{R}_j(\boldsymbol{x}) = |R_j(\boldsymbol{x}) - aR_{j-4}(\boldsymbol{x})|^+, \quad 5 \le j \le 8, \end{aligned} \tag{5}$$

where $\boldsymbol{x} = (x, y, t)$ is a 3D position in the space-time, a controls the weight of opposite motion and is set to 1 here, and $|\cdot|^+$ is defined as $|z|^+ = \max(0, z)$.

motion	Flicker	Left	Left-Up
filter	g_0^{3d}:(0,0,1)	g_1^{3d}:$(\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2})$	g_2^{3d}:$(\frac{1}{2}, -\frac{1}{2}, \frac{\sqrt{2}}{2})$
motion	Up	Right-Up	Right
filter	g_3^{3d}:$(0, -\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$	g_4^{3d}:$(-\frac{1}{2}, -\frac{1}{2}, \frac{\sqrt{2}}{2})$	g_5^{3d}:$(-\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2})$
motion	Right-Down	Down	Left-Down
filter	g_6^{3d}:$(-\frac{1}{2}, \frac{1}{2}, \frac{\sqrt{2}}{2})$	g_7^{3d}:$(0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$	g_8^{3d}:$(\frac{1}{2}, \frac{1}{2}, \frac{\sqrt{2}}{2})$

Table 1. The nine 3D Gabor filters with different directions and their corresponding motions. For example, the 3D Gabor filter g_1^{3d} with $(\eta_x, \eta_y, \eta_z) = (\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2})$ is sensitive to motion towards left.

The motion salience E is measured by the summation of all the oriented motion energies

$$E(\boldsymbol{x}) = \sum_{k=0}^{8} \bar{R}_k(\boldsymbol{x}). \tag{6}$$

A threshold ϵ_s is used to detect region with high motion saliency and generate a binary motion salient region (MSR)

$$S(\boldsymbol{x}) = \begin{cases} 1 & \text{if } E(\boldsymbol{x}) > \epsilon_s, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The definition of MSR involves all the oriented motion energies. In order to emphasize motion along a specific direction, we define the oriented motion salient region (OMSR) as

$$S_k(\boldsymbol{x}) = \begin{cases} 1 & \text{if } S(\boldsymbol{x}) = 1 \text{ and } \bar{R}_k(\boldsymbol{x}) > \epsilon_k, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

where ϵ_k is a threshold and $0 \le k \le 8$. The OMSRs $\{S_k\}_{k=0}^{8}$ decompose S, but there may be overlaps between different OMSRs.

4 Oriented Motion Salient Descriptors

Having detected a set of OMSRs for a video sequence, we construct a low-level OMSD to describe each OMSR. To compute an OMSD, we first extract the texture information of each pixel in the OMSR. We then create a texture polar histogram to describe the salient region in each frame and combine all the polar histograms from the video sequence together to form the final OMSD.

The texture information of the pixels in the OMSR is captured by a bank of 2D Gabor filters. A 2D Gabor filter consists of the real part and imaginary part, which are defined as

$$g_r^{2d}(x,y) = \exp(-\frac{x'^2 + \gamma y'^2}{2\sigma^2})\cos(\frac{2\pi}{\lambda}x') \tag{9}$$

and

$$g_j^{2d}(x,y) = \exp(-\frac{x'^2 + \gamma y'^2}{2\sigma^2})\sin(\frac{2\pi}{\lambda}x') \tag{10}$$

respectively, where $x' = x\ \cos\theta + y\ \sin\theta$, and $y' = -x\ \sin\theta + y\ \cos\theta$, θ controls the direction of the 2D Gabor filter, and γ is spatial respect ratio. An overall response of a 2D Gabor filter is obtained by squaring and summing the outputs of the two part filters. There are 5 scales: $\sigma \in \{5, 7, 9, 11, 13\}$ and 6 directions: $\theta \in \{0°,\ 30°,\ 60°,\ 90°,\ 120°,\ 150°\}$ in the bank. There are in total 30 filters in this bank. Through convoluting the video sequence with all the filters, a 30 dimensional response vector is obtained for each pixel.

For the tth frame, a polar coordinate system is applied to model the spatial distribution of the salient regions in this frame. The origin of the polar coordinate system is set as the geometric center l^t of the salient region of MSR S in the tth frame. We divide the polar coordinate system into N_1 cells. For each cell, we build a histogram of the 2D Gabor filter responses at different orientations and scales. The histogram is computed as a summation of response vectors of pixels in the region belonging to this cell. The final polar histogram for the whole salience region in this frame is computed as a concatenation of the histograms from all the cells. There are total $30{*}N_1$ bins in the polar histogram. So, for this frame, we obtain (h_k^t, l^t), where h_k^t denotes obtained polar histogram, and $l^t = (x_t, y_t)$ is the geometric center. Figure 2 shows the process of the construction of the texture polar histogram in a frame.

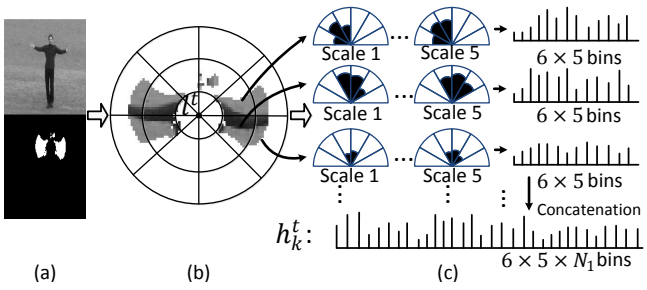

Fig. 2. (a) The tth frame of a video sequence and its one OMSD S_k in this frame. (b) A polar coordinate system. (c) The construction of the polar histogram in this frame.

The global representation for OMSR S_k, is computed as a summation of all the h_k^t from the sequence, taking into acount the spatial distribution of the l^t

in each frame. Similarly, we apply another polar coordinate system to describe the relative positions of all the geometric centers $\{l^1, l^2, ...\}$ in all the frames. The origin of the polar coordinate is set as the mean of $\{l^1, l^2, ...\}$. The polar coordinate is used to divided the plane into N_2 cells. We sum up the h_k^t through the sequence whose center point l^t is in the ith cell to generate a global vector

$$H_{k\_i} = \sum_t h_k^t \ \delta(i, cell(l^t)) \tag{11}$$

where $\delta(\cdot,\cdot)$ is a Dirac kernel and $cell(l^t)$ returns the index of the cell where l^t is located. The OMSD is expressed as the concatenation of H_{k_i} from all the cells

$$H_k = [H_{k\_1}^T, H_{k\_2}^T, ..., H_{k\_N_2}^T]^T,$$

with a dimension of $30{*}N_1 * N_2$. Figure 3 gives a illustration of how to combine all the polar histograms from the video sequence into the OMSD.

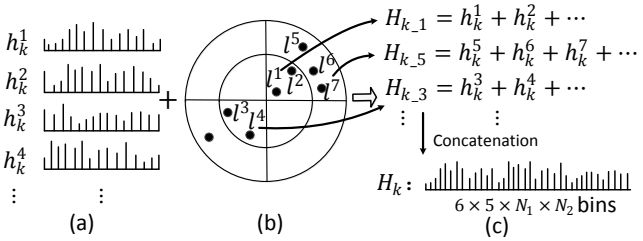

Fig. 3. The process of combining all the polar histograms $\{h_k^1, h_k^2, ...\}$ into OMSD H_k.

The obtained OMSD H_k can be viewed as a two-layered texture polar histogram. The texture information of pixels and spatial distribution information of the OMSR are both included in this descriptor. In fact OMSDs fuse both appearance and motion information in a way quite different from other descriptors, i.e., though encoding only texture information, each OMSD itself corresponds to a specific motion direction. The OMSDs $\{H_k\}_{k=0}^8$ give an effective and informative representation for an action, based on the main motion directions in the action.

5 Class-Specific Oriented Attributes

For an action, we have extracted a set of low-level OMSDs, which describe the action with respect to different motion directions. A traditional way to deal with these OMSDs is to concatenate them together to form a single complex descriptor for action representation. Then the human action recognition problem is transformed into a problem of learning a classification function to assign an action class label to the complex action descriptor. However, actions are so complex that the low-level action descriptors and the action labels cannot capture

all the intrinsic characteristics of actions. We do not use these obtained OMSDs directly for action classification, but instead use them to obtain the class-specific oriented attributes (CSOAs) of each action class. A CSOA describe a specific action class in a specific motion direction, and it is only shared by the actions from this class. So the CSOAs can be used to infer action class labels.

A one-versus-all attribute classifier is trained for each CSOA. Given N training action examples from M action classes, we denote them as $\{(\{H_k^i\}_{k=0}^8, y_i)\}_{i=1}^N$, where $\{H_k^i\}_{k=0}^8$ are the OMSDs for the ith action and $y_i \in \{1, 2, ..., M\}$ is the action class label. Let Φ_y^k denote the attribute classifier for the yth action class on the kth motion direction. The positive examples for this classifier are a set of OMSD $\{H_k^i \mid y_i = y\}$ and the negative ones are $\{H_k^i \mid y_i \neq y\}$. The applied attribute classifier is a simple SVM classifier with a χ^2-kernel to measure the similarity between two OMSDs H_k^i and H_k^j

$$k(H_k^i, H_k^j) = \exp\{-\frac{1}{2\sigma^2}\sum_r \frac{[(H_k^i)_r - (H_k^j)_r]^2}{[(H_k^i)_r + (H_k^j)_r]}\}, \tag{12}$$

where $(H_k^i)_r$ is the rth element of H_k^i and σ is the scale of the kernel function. For the ith action with OMSDs $\{H_k^i\}_{k=0}^8$, the output of attribute classifier Φ_y^k is denoted as $\Phi_y^k(H_k^i)$, which indicates the confidence of the ith action with respect to the kth oriented attribute of the yth action class. Nine attribute classifiers are learnt for each action class and there are $9 * M$ attribute classifiers in total.

The CSOAs are used for action representation because they are effective for distinguishing between actions. For an action x with OMSDs $\{H_k\}_{k=0}^8$, mapping it to the CSOA space of action class y will generate a compact attribute feature vector with respect to class y. The elements of this vector are the outputs of oriented attribute classifiers related to this action class

$$\Psi(x, y) = [\Phi_y^0(H_0), \Phi_y^1(H_1), ..., \Phi_y^8(H_8)]. \tag{13}$$

The attribute feature vector $\Psi(x, y)$ is a middle-level representation for the action. It connects the low-level feature OMSDs $\{H_k\}_{k=0}^8$ and the action class label y together and indicates the confidence of action x belonging to class y with respects to multiple motion directions.

6 Action Classification Based on CSOAs

We apply the learned CSOAs for human action classification. CSOAs characterize each action class with respect to the different directions of the actions. If an action belongs to a certain action class, it should match the CSOAs related to this class well. In this way, the action classification problem is solved by finding the action class whose SCOAs match the given action best.

Given an action, we can obtain a class-specific attribute vector with the form of equation (13) by mapping its low-level OMSDs to the CSOA space of a specific action class. Intuitively, the class-specific attribute vector itself suggests

whether an given action belongs to the specific action class, i.e., if all the entries of the vector have great values, it belongs to this class and vice versa. However, in practice only one or several oriented attributes are of great significance for actions. For example, in videos of "running left", the detected OMSRs are mainly focusing on the directions of left, left-up, left-down and flicker. It means that only the oriented attributes of the above mentioned directions have great discriminative power for recognizing the action "running left" while attributes of other directions have little relevance.

In such case, we define a compatibility function f to measure how well the given action x matches the CSOAs of action class y. We assume that f is a linear combination of the attribute feature vector

$$f(x, y) = \omega_y^T \Psi(x, y), \tag{14}$$

where ω_y is the parameter vector of f associated with class y. It emphasizes the importance of the oriented attributes of a specific class. The function f plays a role as a action classifier and the predicted label y^* for the action x is derived by maximizing f over all $y \in Y$

$$y^* = \arg\max_{y \in Y} f(x, y). \tag{15}$$

The compatibility function f can be learned in an SVM formulation. Given N training examples $\{(x_i, y_i)\}_{i=1}^N$ where x_i is an input video sequence and $y_i \in Y = \{1, 2, ..., M\}$ is the corresponding action label, the parameter vector w_y is learned by solving a convex quadratic optimization problem which is expressed as

$$\min_{\omega_y} \frac{\gamma}{2} \parallel \omega_y \parallel^2 + \sum_{y_i = y} \xi_{1,i} + \sum_{y_j \neq y} \xi_{2,j} + \sum_{y_i = y, y_j \neq y} \xi_{3,ij} \tag{16}$$

$$\text{s.t. } \forall y_i = y, \quad \omega_y^T \Psi(x_i, y_i) \geq 1 - \xi_{1,i},\ \xi_{1,i} \geq 0, \tag{17}$$

$$\forall y_j \neq y, \quad \omega_y^T \Psi(x_j, y_i) \leq -1 + \xi_{2,j},\ \xi_{2,j} \geq 0, \tag{18}$$

$$\forall y_i = y, y_j \neq y, \quad \omega_y^T \Psi(x_i, y_j) \leq -1 + \xi_{3,ij},\ \xi_{3,ij} \geq 0, \tag{19}$$

where $\xi_{1,i}$,$\xi_{2,j}$,and $\xi_{3,ij}$ are the slack variables and γ is a constant that controls the trade-off between training error minimization and margin maximization.

We analyze the above constrains for a better understanding of the learning formulation. For class y, constraint (17) requires that in the training stage, only the attribute vectors generated by mapping the input videos of class y to the CSOAs of class y are used as positive examples while constrains (18) and (19) indicates that when mapping the videos of class y into the CSOAs of other classes or mapping the videos of other classes into CSOAs of class y, the generated attribute vectors are regarded as negative examples. The whole process of learning the function f is quite similar to a multiclass SVM formulation. However, the input attribute vectors for each action class are different because they are determined not only by the low-level features, but also by the action class labels.

7 Experimental Results

We perform a set of experiments to evaluate the performance of our proposed approach on four publicly available datasets: the KTH [15], UCF sports [16], UCF films [16] and Hollywood2 [1]. Some frames extracted from the four datasets are shown in Figure 4.

We conduct two groups of experiments on these datasets. In the first group, we evaluate the performance of our proposed low-level features OMSDs. The extracted OMSDs are directly used for human action representation and an SVM classifier is trained for action classification. In the second group, we evaluate the performance of our CSOAs based approach. After extracting the low-level features OMSDs, we obtain the CSOAs related to each action class and a compatibility classifier is applied for action classification. In addition, we also compare the performance of our CSOAs based approach with some state-of-the-art approaches.

Fig. 4. Some frames are extracted from the four datasets. The rows from top to down are the frames from the KTH dataset, UCF sports dataset, UCF films dataset, and Hollywood2 dataset respectively.

7.1 Datsets and Evaluation Protocol

The KTH dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios. There are in total 599 sequences on this dataset. We perform leave-one-person-out cross validation to make the performance evaluation. In each run, 24 actors' video sequences are used for training and the remaining actor's videos for test.

The UCF sports dataset contains ten sports actions: diving, golf swinging, kicking, lifting, horse riding, running, skateboarding, swinging bench, swinging from side angle and walking. It consists of 150 video sequences taken from actual sporting activities from a variety of sources with a wide range of viewpoints

and scene backgrounds. Leave-one-out cross validation is used to evaluate our approach. One video sequence is used for test and the remaining sequences are used for training.

The UCF feature films dataset provides a representative pool of nature samples of two action classes including kissing and hitting/slapping. It contains 92 samples of kissing and 112 samples of hitting/slapping which are extracted from a range of classic movies. The actions are captured in a wide range of scenes under different viewpoints with different camera movement patterns. The test for this dataset uses leave-one-out cross validation.

The Hollywood2 dataset contains 12 action classes collected from 69 different Hollywood movies. There are in total 1707 video sequences on this dataset, which are divided into a train set of 823 sequences and a test set of 884 sequences. We follow the standard evaluation protocol on this benchmark, i.e., computing the average precision (AP) for each class and using the mean of APs (mAP) for performance evaluation.

7.2 Evaluation of Low-Level OMSDs

We test our low-level feature OMSDs on both the KTH dataset and UCF sports dataset. Given a video sequence containing an action x, we detect the OMSRs and compute the corresponding OMSDs $\{H_k\}_{k=0}^{8}$. In our experiments, we first evaluate the performance of each OMSD. The whole video sequence is represented by only a single OMSD H_k. A simple SVM classifier with χ^2-kernel is applied for human action classification. Then we utilize a feature-level fusion approach in which all the OMSDs $\{H_k\}_{k=0}^{8}$ are concatenated together to form a large feature vector

$$H = (H_0^T, H_1^T, ..., H_8^T)^T \tag{20}$$

for action representation. Similarly, this large feature vector is supplied as input to a SVM with χ^2-kernel for action classification. The performance of each OMSD and the concatenation of OMSDs on both datasets are shown in Table 2. In Table 2, 'OMSD-k' means that video sequences are represented by the kth OMSD H_k, and 'Average of OMSDs' is computed as the mean of accuracies of the 9 OMSDs.

Three points can be drawn from Table 2. First using only a single low-level feature OMSD obtains relative good results, varying from 91.0% to 95.6% on the KTH dataset and from 80.7% to 85.7% on the UCF sports dataset. This demonstrates the effectiveness of the proposed OMSDs. Since the construction of a OMSD is based on a specific OMSR detected in a video sequence, it captures only a small part of the visual information of an action. However, the average accuracy of the OMSD still reaches 92.8% on the KTH dataset and 82.4% on the UCF sports dataset.

Second, the large feature H obtained by concatenating all the OMSDs achieves the best results on both datasets, reaching 96.5% for the KTH dataset and 88.0% for the UCF sports dataset. This is about 3.7% and 5.6% higher than the average accuracy of single OMSD on the KTH and UCF sports datasets respectively. We

Table 2. Comparison the performance of each low-level feature OMSD and the concatenation of OMSDs on the KTH and UCF sports datasets.

Low-Level Features	KTH	UCF sports
OMSD-0	**95.6%**	**85.3%**
OMSD-1	91.0%	82.6%
OMSD-2	91.2%	83.3%
OMSD-3	94.0%	81.3%
OMSD-4	93.5%	80.7%
OMSD-5	93.3%	80.7%
OMSD-6	92.1%	84.0%
OMSD-7	91.6%	82.6%
OMSD-8	92.6%	81.3%
Average of OMSDs	92.8%	82.4%
Concatenation of OMSDs	**96.5%**	**88.0%**

can see that the simple concatenation of all the OMSDs can improve the performance of human action recognition by a large amount.

Third, when using a single OMSD for action representation, the 'OMSD-0' outperforms other 'OMSD-k's, reaching accuracies of 95.6% and 85.3% respectively on the two datasets. In our experiments, the 'OMSD-0' corresponded to the 3D Gabor filter with $(\eta_x, \eta_y, \eta_t) = (0, 0, 1)$. Except for the Gaussian scales, this 3D Gabor filter is equivalent to Dollár's linear separable filters designed for spatiotemporal interest points detection. This 3D Gabor filter generates a large response where motion occurs, regardless of the motion direction.

7.3 Evaluation of Middle-Level CSOAs

In this subsection, we evaluate the performance of our CSOAs based approach for human action recognition on the four datasets mentioned above. We extract the low-level OMSDs for each video sequence and utilize these OMSDs to train the CSOA classifiers for each action class. Then mapping actions into the CSOA space of each action class, a set of middle-level attribute features are constructed, which combine the low-level OMSDs and action class labels together. A compatibility function is learned to measure how well the low-level features match the CSOAs of action classes.

Table 3 presents a comparison of our proposed CSOA based approach with other approaches on the KTH and UCF sports datasets. Our CSOA based approach outperforms the other methods, achieving 97.2% and 91.3% on these two datasets respectively, which demonstrates the effectiveness of our CSOAs based approach. It is notable that the performance of our CSOA based approach is 0.7% higher than that of the low-level concatenation of OMSDs on the KTH dataset and 3.3% on the UCF sport dataset. This shows that firstly the CSOAs learnt from low-level OMSDs carry great discriminative power and improve the performance of human action recognition and secondly only the concatenation

Table 3. Comparison of our CSOA based approach with state-of-the-art approaches on the KTH and UCF sports datasets.

Algorithm	KTH	UCF sports
Derpanis *et al.* [17]	93.2%	81.5%
Wang *et al.* [4]	92.1%	85.6%
Kovashka *et al.* [18]	94.5%	87.3%
Le *et al.* [19]	93.9%	86.5%
Wang *et al.* [20]	94.2%	88.2%
Liu *et al.* [25]	94.8%	-
Wang*et al.* [23]	93.3%	-
Shi *et al.* [24]	93.0%	-
Concatenation of OMSDs	96.5%	88.0%
Our CSOAs	**97.2%**	**91.3%**

Table 4. Confusion table of our CSOA based approach on the KTH dataset.

	Box	Handclap	Handwave	Jog	Run	Walk
Box	**1.00**					
Handclap	0.01	**0.98**	0.01			
Handwave		0.01	**0.99**			
Jog				**0.95**	0.03	0.02
Run				0.08	**0.92**	
Walk				0.01		**0.99**

Table 5. Confusion table of our CSOA based approach on the UCF sprots dataset.

	Dive	Golf	Kick	Lift	Ride	Run	Skate	Swing1	Swing2	Walk
Dive	**1.00**									
Golf		**0.95**				0.05				
Kick			**1**							
Lift				**1**						
Ride		0.08			**0.83**	0.08				
Run		0.08			0.15	**0.69**		0.08		
Skate		0.16			0.08		**0.75**			
Swing1					0.05			**0.95**		
Swing2									**1.00**	
Walk						0.09				**0.91**

Table 6. The results of our approach on the UCF film dataset.

Algorithms	Kiss	Slap	Average
Rodriguez *et al.* [16]	66.4%	67.2%	66.8%
Yeffet *et al.* [21]	77.3%	84.2%	80.75%
Concatenation of OMSDs	92.2%	93.2%	92.7%
Our CSOAs	**95.6%**	**96.5%**	**96.1%**

Table 7. The results of our approach on the Hollywood2 dataset.

Algorithms	mAP
Wang *et al.* [4]	47.7%
Le *et al.* [19]	53.3%
Wang *et al.* [20]	58.3%
Concatenation of OMSDs	52.6%
Our CSOAs	**58.6**%

of low-level OMSDs achieves a good performance on the KTH dataset, because the actions on the KTH dataset are simple actions performed against static and un-cluttered backgrounds. The confusion matrixes of the proposed CSOA based approach on the KTH and UCF sports datasets are shown in Table 4 and Table 5 respectively.

Table 6 and Table 7 show the performance of our approach on both UCF films and Hollywood2 datasets. Our CSOAs based approach achieves 96.1% and 58.6% respectively on both datasets which is comparable to the listed approaches. It demonstrates the effectiveness of our CSOAs based approach on the realistic datasets. Meanwhile, the CSOAs based approach is 3.4% and 6.0% respectively higher than the simple concatenation of all low-level OMSDs. It indicates the CSOA based approach outperforms the low-level OMSDs based approach.

8 Conclusion

In this paper, we have proposed a novel approach for human action recognition based on the oriented motion salient regions. First, a 3D Gabor filter bank, incorporated with an opponent inhibition operator, has been applied to detect the OMSRs and a set of OMSDs have been extracted from these detected regions. Then, the obtained OMSDs have been used to explore the oriented characteristics of each action class, obtaining a series of CSOAs for each class. Taking advantage of these CSOAs, we have obtained a compact and discriminative middle-level feature to represent human actions. Finally, a compatibility function has been devised for action classification. We have tested our proposed approach on several public datasets. The experimental results have demonstrated that the proposed approach are effective in human action recognition.

Acknowledgement. This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), the Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081) and NSFC (Grant No. 61100099, 61303086).

References

1. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009) 2929–2936
2. Dollár, P., Rabaud, V., Cottrell, Ga., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VSPTES (2005) 65–72
3. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008) 1–8
4. Wang, H., Ullah, M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In:BMVC (2009)
5. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional SIFT descriptor and its application to action recognition. In: ICMM (2007) 357–360
6. Bobick, A.F. and Davis, J.W.: The recognition of human movement using temporal templates. PAMI **23** (2001) 257–267
7. Wang, L., Suter, D.: Informative shape representations for human action recognition. In: ICPR Volume 2 (2006) 1266–1269
8. Ikizler, N., Duygulu, P.: Histogram of oriented rectangles: A new pose descriptor for human action recognition. IVC **27** (2009) 1515–1526
9. Reed, Todd, R.: Motion analysis using the 3-d gabor transform. SSC **1** (1996) 506–509
10. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A **2** (1985) 284–299
11. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009) 1778–1785
12. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011) 3337–3344
13. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: ECCV (2010) 155–168
14. Yao, B., Jiang, X., Khosla, A., Lin, A., L., Guibas, L., Fei-Fei, L.:, Human action recognition by learning bases of action attributes and parts. In: ICCV (2011) 1331–1338
15. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR **3** (2004) 32–36
16. Rodriguez, M., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR (2008) 1–8
17. Derpanis, K., Sizintsev, M., Cannons, K., Wildes, R: Action Spotting and Recognition Based on a Spatiotemporal Orientation Analysis. PAMI **35** (2012) 527–540
18. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR (2010) 2046–205
19. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR (2011) 3361–3368
20. Wang, H., Klaser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: CVPR (2011) 3169–3176
21. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: ICCV (2009) 492–497
22. Derpanis, K., Lecce, M., Daniilidis, K., Wildes, R.P.: Dynamic scene understanding: The role of orientation features in space and time in scene classification. In: CVPR (2012) 1306–1313

23. Wang, L., Qiao, Y., Tang, X.: Motionlets: Mid-Level 3D Parts for Human Motion Recognition. In: CVPR (2013)
24. Shi, F., Petriu, E., Laganiere, R.: Sampling strategies for real-time action recognition. In: CVPR (2013) 2595–2602
25. Liu, L., Shao, L., Zhen, X., Li, X.: Learning Discriminative Key Poses for Action Recognition. Cybernetics (2013)