

## Accepted Manuscript

Distinctive action sketch for human action recognition

Ying Zheng, Hongxun Yao, Xiaoshuai Sun, Sicheng Zhao,  
Fatih Porikli

PII: S0165-1684(17)30381-X  
DOI: [10.1016/j.sigpro.2017.10.022](https://doi.org/10.1016/j.sigpro.2017.10.022)  
Reference: SIGPRO 6643

To appear in: *Signal Processing*

Received date: 16 March 2017  
Revised date: 16 October 2017  
Accepted date: 19 October 2017

Please cite this article as: Ying Zheng, Hongxun Yao, Xiaoshuai Sun, Sicheng Zhao, Fatih Porikli, Distinctive action sketch for human action recognition, *Signal Processing* (2017), doi: [10.1016/j.sigpro.2017.10.022](https://doi.org/10.1016/j.sigpro.2017.10.022)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



**Highlights**

- We introduce sketch to the field of action recognition and propose a ranking based method to discover the most distinctive action sketches.
- We present a novel approach of action representation based on four kinds of sketch pooling strategy.
- Extensive experiments on two public human action datasets are conducted to demonstrate the effectiveness of the proposed method.

ACCEPTED MANUSCRIPT

## Distinctive action sketch for human action recognition

Ying Zheng<sup>1,2</sup>, Hongxun Yao<sup>1</sup>, Xiaoshuai Sun<sup>1</sup>, Sicheng Zhao<sup>3</sup>, Fatih Porikli<sup>2</sup>

<sup>1</sup> School of Computer Science and Technology, Harbin Institute of Technology, China

<sup>2</sup> Research School of Engineering, Australian National University, Australia

<sup>3</sup> School of Software, Tsinghua University, China

{zhengying, h.yao, xiaoshuaisun}@hit.edu.cn, schzhao@gmail.com, fatih.porikli@anu.edu.au

---

### Abstract

Recent developments in the field of computer vision have led to a renewed interest in sketch correlated research. There have emerged considerable solid evidence which revealed the significance of sketch. However, there have been few profound discussions on sketch based action analysis so far. In this paper, we propose an approach to discover the most distinctive sketches for action recognition. The action sketches should satisfy two characteristics: sketchability and objectiveness. Primitive sketches are prepared according to the structured forests based fast edge detection. Meanwhile, we take advantage of Faster R-CNN to detect the persons in parallel. On completion of the two stages, the process of distinctive action sketch mining is carried out. After that, we present four kinds of sketch pooling methods to get a uniform representation for action videos. The experimental results show that the proposed method achieves impressive performance against several compared methods on two public datasets.

*Keywords:* Action Sketch, Sketch Pooling, Action Recognition

---

## 1. Introduction

With the flourish of computer vision, sketch based technology is becoming a rising star for multitudinous researchers all over the world. It is well acknowledged that sketch has an important practical significance. In the past few years, there have been numerous works tackling sketch related problems from different angles. They mainly focused on three areas, namely sketch based image retrieval [1, 2] and composition [3, 4], sketch based video retrieval [5, 6], sketch segmentation [7] [8] and recognition [9, 10]. Although it is a widespread idea since the success of sketch based approaches that studying the sketch of visual elements is one of the most fundamental prerequisites for many vision applications, there is still limited studies revolving around action sketch.

We can probably find the trace of action sketch in the system of sketch based video retrieval. But almost all of existing works concentrate on how to create a better algorithm for video clip searching, departed from free-hand sketch queries which depict the shape, color and movement of objects roughly [5]. Unlike any of these sketch based video retrieval works, we do not seek to achieve determinate mapping between input sketch and video. In fact, our problem goes retrograde in some ways, because we want to transform action video into sketch. The progress of this issue be a powerful auxiliary tool for many works, such as action recognition, and retrieval [11] [12].

Among these scarce groups associated with action sketch, A.Yilmaz's work [13] is the most similar one to ours. They present a method of action representation on spatio-temporal volume (STV) and differential geometric surface properties. How to represent action better for tasks like action recognition is what they chased. Besides, the first premise they assumed is that object contours for each action slice are given. In contrast, our work principally seeks conversion between action and sketch which is one step ahead of action representation.

In this paper, we propose a method of distinctive action sketch mining for human action recognition. First of all, we generally explore the characteristic

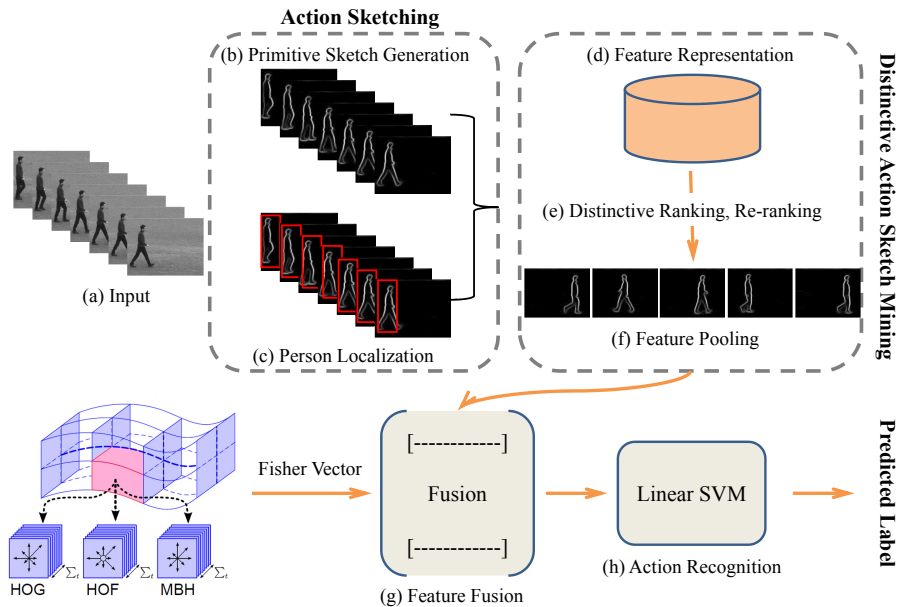


Figure 1: Overview of our method. (a) Input: action videos. (b) Transform each action frame to primitive sketch in real time by fast edge detection method [14]. (c) Locate person by Faster R-CNN [15]. (d) Represent sketches as feature vectors. (e) Discover the top distinctive action sketches through ranking and re-ranking. (f) Apply feature pooling on the top ranking sketches to get a new representation. (g) Perform representation-level fusion with improved dense trajectories with Fisher vector encoding. (h) Recognize action videos by a linear SVM and choose the prediction with highest score as the predicted label.

of sketch in action and build an applicable system to discover the most distinctive action sketches possessing sketchability and objectiveness. For action videos, sketches of each clip can be well generated in real time. Combining these elaborate sketches, we propose a distinctive ranking method of action sketches.

<sup>35</sup> The top ranking sketches can typically represent the action classes which they belong to. Among the obtained top ranking sketches, we introduce an approach of feature pooling to get a new representation for action video. Then the new representation will be combined with local feature based representation such as the improved dense trajectories with Fisher vector encoding. Figure 1 gives the

<sup>40</sup> overview of our method.

The main contributions of this work are summarized in three-fold:

- We introduce sketch to the field of action recognition and propose a ranking based method to discover the most distinctive action sketches.
- We present a novel approach of action representation based on four kinds of sketch pooling strategy.
- Extensive experiments on two public human action datasets are conducted to demonstrate the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section 2 reviews the works related to our research. Section 3 shows how to transform the human action to sketches. Section 4 describes the distinctive ranking method of action sketches. Section 5 introduces the approach of feature pooling for action recognition. Section 6 presents the experimental results, followed by conclusion in Section 7.

## 2. Related work

As one of the most important research fields in computer vision, human action recognition has a wide range of applications, such as human-computer interaction, video surveillance and robot action control. In the past years, numerous approaches have been proposed to understand and recognize human actions from different angles and levels. Among these works, action representation is a key step towards a good action recognition system [16]. In practice, a human action clip can be represented by different views or features, such as motion, gradients, and shapes. The action sketch we proposed can be seen as one kind of views. In this section, we will give a brief introduction of feature representation for action recognition and several other works closely related to our research. For a wider range of studies on action recognition, we recommend the insightful reviews [11] [17] [18] to interested readers.

Traditional approaches for action recognition are mainly based on single feature representation, such as local and global feature representation [19] [20]

[21]. Based on hand-crafted descriptors or neural networks, these methods can  
70 achieve good performance. However, the problems of lighting and viewpoint  
changes, complex backgrounds and intra-class variations have made it very chal-  
lenging to get a higher accuracy. To address this problem, many researchers  
propose new representations which combine different features together. In con-  
sideration of the particular characteristics possessed by different kind of features,  
75 some well-designed combinations are usually superior to the single feature rep-  
resentation. These methods can be classified into two groups, direct catenation  
[22] and multi-view learning [23] [24] [25]. For direct catenation, space-time  
interest points (STIP) [26] and improved dense trajectories (IDT) [27] are the  
most classic works. Under the standard bag of visual words (BoVW) frame-  
80 work, Laptev et al. [28] demonstrate the effectiveness of feature combination of  
the histograms of oriented gradient (HOG) descriptors and histograms of opti-  
cal flow (HOF) descriptors computed for STIP. For IDT based representation,  
Wang et al. [27] catenate four descriptors (i.e., trajectory, HOG, HOF and  
motion boundary histograms (MBH)) coded by Fisher vector and obtain the  
85 state-of-the-art performance for action recognition at that time. Different from  
direct catenation, multi-view learning based methods focus on exploring the re-  
lationship between different features [29] and incorporate these heterogeneous  
feature descriptors into one low-dimensional and compact representation [30].  
Although this two kinds of works have explored numerous and diverse features,  
90 taking the sketch to represent action is a new and largely unexploited frontier.

An early preliminary version of this work was published in [31]. Compared  
to the earlier version, the biggest difference is that we propose a novel method  
of action representation based on sketch pooling. We also replace the person  
detection method with the state-of-the-art Faster R-CNN [15]. Furthermore,  
95 we present a unified approach for distinctive action sketch mining. Among  
other related works, silhouette and skeleton are similar to action sketch. The  
silhouette is usually a black-white image using the edges to draw the outline of  
person or object. It is a critical research medium in the field of human pose  
recovery. Typical approaches of image-based pose recovery reconstruct 3D poses

100 by the learned mapping function between 2D silhouettes and 3D poses [32] [33]  
[34]. For human skeleton, it is composed of several rigid segments which are  
articulated by joints. The skeleton looks like a matchstick man that forms the  
rough structure of a human body. By modeling and classifying the temporal  
evolution of human skeleton, human action can be recognized in a 3D feature  
105 space [35] [36]. Unfortunately, both silhouette and skeleton are mostly applied  
in 3D video sequences that require a sophisticated motion capture system or  
effective depth sensors. For this reason, they cannot be implemented to the  
conventional human action videos while the proposed action sketch does not  
have this limitation.

### 110 3. Action sketching

When we talk about painting art, a sketch usually refers to a quick and  
informal drawing done from real life. An excellent sketch should capture the  
essentials of a subject, which may be the overall neurogram and characteristic  
details from a specific perspective. In a general sense, the only primary mission  
115 for static image sketching is to hit the spot of greatest possible similarity of both  
dominant lines. However, moving to action videos, the situation is quite different  
in the aspect of object's subjectivity. Meanwhile, it is the key factor we need  
to consider. We implement appropriate measures to satisfy the requirement in  
action sketching.

#### 120 3.1. Primitive sketch generation

In the previous works, Yilmaz et al. take the object contours as their basic  
elements [13]. Admittedly, object contours may have some certain degree of  
ability to express the sketch of action. Nonetheless, the ability is not enough to  
fully appear in the person of action sketch. Beyond that, methods like tracka-  
125 bility maps [37] and action templates [38] are also proposed to represent action  
videos. Leaving aside the performances of these methods, the outcome has  
identifiable differences as compared with the sketch.



In our research, the primary character for sketching action will be referred to as *sketchability*. Methods that meet the requirements of sketchability must be able to depict the full profile of subject similar to object contours and some other important details. It should be pointed out that the transformation between real image and sketch is immensely challenging [39]. This is not our work in this paper. A reasonable way is to select some appropriate representative edges for each action clip. To generate the primitive sketch, we adopt the fast edge detection method proposed by P. Dollár et al. [14]. In consideration of the structure underlie local image patches, they presented a structured learning framework for local edge mask prediction integrated into random decision forests. In this stage, we can transform each action frame to primitive sketch in real time.

### 3.2. Person localization

What we are talking about here is human-centered action videos. Of course, it can be easily extended to other types of subjects. The most important characteristic of these videos is that the actions must be performed by a specific subject. We call this property of action videos *objectiveness*.

To capture the objectiveness in action videos, we use the method called Faster R-CNN [15] as an accurate and efficient tool for person localization. The Faster R-CNN depends on region proposal network (RPN) that shares full-image convolutional features with the detection network, which makes it become the state-of-the-art object detection network. In practice, we use the open sourced code of Ren et al. [15] and the ImageNet pre-trained network released on their website<sup>1</sup> to detect the person in each frame. Then we filter out the detections with too small size and unsatisfied aspect ratio for better model detection. It is necessary to note that this stage is completely independent of the stage of primitive sketch generation, so we can deal with this two processes in two parallel channels, which can provide significant run time savings. After that, we will get

<sup>1</sup>[https://github.com/ShaoqingRen/faster\\_rcnn](https://github.com/ShaoqingRen/faster_rcnn)

a corresponding sketch which is only associated with the subject for every action clip when combining these two stages together.

#### 4. Distinctive action sketch mining

It is generally known that there are discriminative and representative themes with semantic interpretation in a specific type of photos, such as city and landscape images [40] [41]. Similarly, some distinctive patterns also exist in different types of actions. While the actions appear in the form of sketch, distinctive sketches will correspond to these patterns. Acquisition of these sketches can be a significant assistant to many action related applications. In this section, we will present the details of the proposed method for distinctive action sketch mining.

##### 4.1. Feature representation

Given a sequence of action sketches generated in the stage of action sketching, we must first find a fine feature representation for every sketch. It is generally agreed that a deeper convolutional neural network can describe richer semantic information of the image. However, the very deep network calls for large-scale labeled training dataset, which is not always available in research. Fortunately, there are a handful of superior deep models for image classification from ImageNet large scale visual recognition challenge (ILSVRC). These deep nets are delicately designed and well trained with millions of images from ImageNet. Ali et al. [42] report that the pretrained deep model from ILSVRC is able to get consistent superior results on a diverse range of tasks including scene recognition, image retrieval, etc. Notably, the recent work of Sketch-a-Net proposed by Yu et al. [43] present persuasive evidence that deep model is a super choice for representation of sketch. To get better representative features, we employ the VGG-19 Net [44] and extract features of each sketch from the fully-connected (FC) layer. Finally, the sketch will be represented by a feature vector  $X$  with 4096 dimensions.

#### 4.2. Distinctive ranking

185 We want to discover these sketches which can distinctively represent a particular category of action. In other words, these distinctive action sketches have to meet requirements of the following two aspects: a) sketches must be very representative in the class they belong to; b) they also should be diverse among the sketches we selected.

For the first requirement, we propose a distinctive ranking method based on clustering. Among these feature vectors  $X$  of sketches, we gather them into  $k$  clusters by  $k$ -means and get the feature vectors  $Y$  of each cluster center. After that, all sketches are ranked according to the distinctive score computed by

$$S = \frac{\sum_{i=1}^k d_{12}(X_c, Y_i)}{d_{12}(X_c, Y_s)}, \quad (1)$$

190 where  $d_{12}$  refers to the Euclidean distance,  $X_c$  and  $Y_s$  are the feature vector of current sketch and the center of its cluster respectively.

In the fraction above,  $d_{12}(X_c, Y_s)$  denotes distance between current sketch and its cluster center. A smaller value means that the sketch is more likely to represent the class it belongs to. On the other hand,  $\sum_{i=1}^k d_{12}(X_c, Y_i)$  indicates the sum of distance from the sketch to every cluster center. A larger value shows that the sketch is more diverse among these sketches. The scores by division between them can trap the degree of distinctive property. What we want is that the most representative sketches have the highest order. It can be represented by a function  $D$  of distinctive score maximization, which is formulated as follows,

$$\begin{aligned} D &= \max_{\theta} \sum_{i=1}^n (n - i + 1) S_{\theta_i} \\ &= \max_{\theta} \left\{ (n + 1) \sum_{i=1}^n S_{\theta_i} - \sum_{i=1}^n i S_{\theta_i} \right\} \end{aligned} \quad (2)$$

where  $\theta$  is an array with  $n$  parameters that indicate the positions to place each sketch,  $n$  means the number of sketches in an action clip, and  $S_{\theta_i}$  denotes the distinctive score of the  $\theta_i$ -th sketch. Only when the value of  $S_{\theta_i}$  is greater than  $S_{\theta_{i+1}}$  ( $i = 1, 2, \dots, n - 1$ ), the equation (2) can get the maximal value. Given

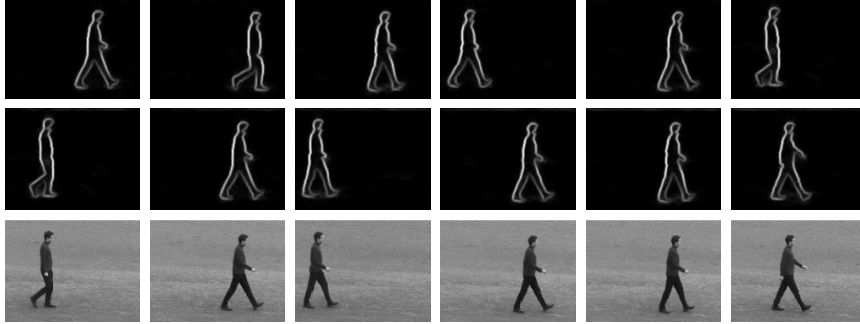


Figure 2: Illustration of distinctive action sketch mining. Top row: top 6 ranking distinctive action sketches. Second row: results after re-ranking. Third row: original frames corresponding to the re-ranking results.

the distinctive scores of all sketches computed by equation (1), the first item of equation (2) equals to a constant  $C$ . Hence the equation is converted to

$$\begin{aligned}
 D &= \max_{\theta} \left\{ C - \sum_{i=1}^n i S_{\theta_i} \right\} \\
 &= \min_{\theta} \sum_{i=1}^n i S_{\theta_i}
 \end{aligned} \tag{3}$$

Now the maximization function of distinctive score becomes a process to find the arrangement of sketches that can minimize the value of equation (3). To achieve this goal, all sketches are ranked in descending order based on the distinctive scores.

#### 4.3. Re-ranking

Distinctive ranking gives us an acceptable mining result of action sketches, but it is coarse somewhat and has a potential shortcoming. From the top row of Figure 2, we can see an undesired phenomenon that several similar sketches have close order. That is to say, the ranked sketches cannot satisfy the second requirement of distinctive action sketch. In such cases, a sectionalized re-ranking approach is presented to overcome the shortcoming.

In consideration of practical necessity and computation cost, we limit the re-ranking process to a separate interval  $w$  of ranking order. When the top

$m$  ranking sketches are the expected output, we do re-ranking among double ranking sketches, i.e. the interval  $w$  equals  $2m$ . For example, if we set  $m$  as 10, then we only need to do re-ranking on the top 20 distinctive action sketches. At the individual range, we iteratively find the most dissimilar one with sorted sketches in the remaining samples. The procedure can be formalized as:

$$\max_p \sum_{k=1}^{j-1} \sqrt{\sum_{t=1}^d (x_{pt} - x_{kt})^2}, \quad (4)$$

where  $p = \{j, \dots, w\}$ ,  $j$  marks the current position calling for re-ranking in the process of iteration,  $d$  means the dimension of feature vector  $X$ . The re-ranking process is designed to maximize the diversity of action sketches obtained in the stage of distinctive ranking. Based on equation (4), it can be written as

$$R = \max_{\varphi} \sum_{j=2}^w \sum_{k=1}^{j-1} d_{12}(X_{\varphi_j}, X_{\varphi_k}), \quad (5)$$

where  $\varphi$  is a new arrangement for the top  $w$  action sketches obtained in the distinctive ranking stage,  $d_{12}(X_{\varphi_j}, X_{\varphi_k})$  computes the Euclidean distance between feature vectors of the  $\varphi_j$ -th and  $\varphi_k$ -th action sketch. By combining equation 3 and (5), the process of distinctive ranking and re-ranking can be unified into a max-min formulation,

$$\max_{\varphi} \min_{\theta} \sum_{i=1}^n iS_{\theta_i} + \sum_{j=2}^w \sum_{k=1}^{j-1} d_{12}(X_{\varphi_j}, X_{\varphi_k}). \quad (6)$$

Figure 2 shows an illustration of distinctive action sketch mining. After the operation of re-ranking on the sketches generated above, the discovered action sketches are more distinctive and acceptable.

## 5. Sketch pooling for action recognition

After obtained the top  $m$  distinctive ranking sketches, the most important task is to get a uniform representation for action videos. The feature vectors of top  $m$  sketches for each video are donated as  $\{X'_1, X'_2, \dots, X'_m\}$ . Based on these

vectors, our goal is to generate a feature vector  $X''$  with the same  $d$  dimensions.

$$X'' = [x''_1, x''_2, \dots, x''_d]. \quad (7)$$

We apply four kinds of feature pooling methods to achieve the task, respectively are average pooling, max pooling, min pooling and every pooling. For multi-class action classification, we implement these methods and want to find the most suitable one by the evaluation on public human action datasets.

- **Average pooling.** The average operation is performed on the front  $u$  feature vectors of top  $m$  sketches which is defined as follows,

$$X'' = \sum_{i=1}^u X'_i / u. \quad (8)$$

In fact, it is taking the mean value of each dimension in the  $u$  vectors as the new element  $x''_t$  of  $X''$ ,

$$x''_t = \sum_{i=1}^u x'_{it} / u, \quad (9)$$

in which  $t = 1, 2, \dots, d$ .

- **Max pooling.** It select the maximum value of the front  $u$  feature vectors which is defined as follows,

$$X'' = \max X'_i, \quad (10)$$

where  $i = 1, 2, \dots, u$ . Max pooling means that the new element  $x''_t$  equals to the maximum value of each dimension,

$$x''_t = \max x'_{it}. \quad (11)$$

- **Min pooling.** On the contrary to max pooling, it chooses the minimum value of each dimension,

$$x''_t = \min x'_{it}. \quad (12)$$

- **Every pooling.** The aim of every pooling is to produce the chance of representing action video by a single sketch. It simply takes the  $u$ -th feature vector as a new representation which is defined as follows,

$$X'' = X'_u. \quad (13)$$

Although the distinctive action sketch can express spatial information very well, it has a big drawback of lacking temporal cues. In order to solve the problem, we propose to combine the pooling feature and improved dense trajectories (IDT) based features. For representation-level fusion, we concatenate the feature vector of action sketch and the normalized Fisher vectors of IDT descriptors into a single one. The obtained feature vector will be further fed into a linear SVM classifier for action recognition.

## 6. Experiments

In this section, we describe the detailed experimental settings and show the results on two public human action datasets. We first introduce the datasets used for evaluation and their corresponding experimental setup. Then we present implementation details of our experiments. After that, we evaluate the performance of our method for action recognition and explore different factors that may impact on the final recognition accuracy.

### 6.1. Datasets

We conduct experiments on two public datasets, respectively are KTH [45] and UCF101 action recognition dataset [46]. The KTH dataset is relatively simple while the UCF101 dataset is more complicated as it is a realistic action dataset collected from YouTube. Some examples of video frames from the two action datasets are illustrated in Figure 3 and Figure 4.

The KTH dataset consists of 600 video files in total and each class has 100 videos which have a uniform resolutions of  $160 \times 120$  pixels<sup>2</sup>. The videos

---

<sup>2</sup><http://www.nada.kth.se/cvap/actions>

are collected from 4 different scenarios and evenly divided into six types of  
 235 actions: walking, jogging, running, boxing, hand waving and hand clapping.  
 We train models on the training + validation set and report average accuracy  
 for evaluation on the test set.

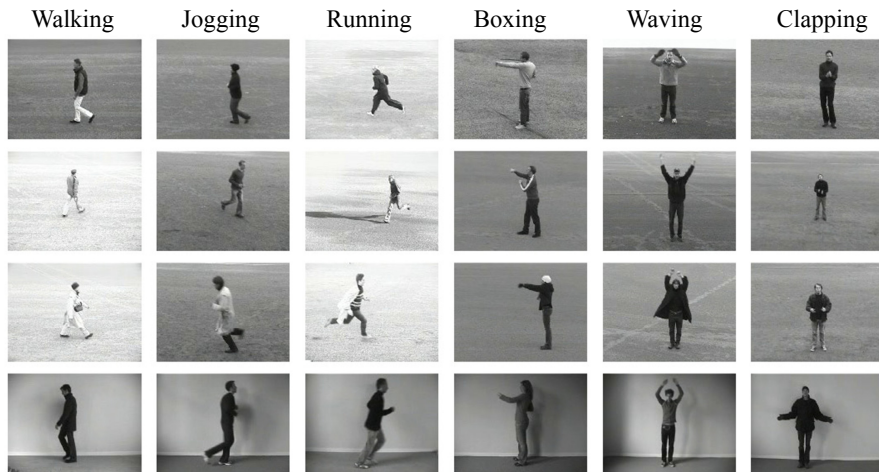


Figure 3: Sample frames from the KTH human action dataset with six classes (columns) and four scenarios (rows) presented.



Figure 4: Sample frames of 8 action classes from the UCF101 dataset. From left to right, the first row: ApplyEyeMakeup, Biking, BlowingCandles, Diving. Second row: Haircut, HorseRiding, PlayingViolin, Surfing.

The UCF101 dataset<sup>3</sup> has 101 action classes and can be divided into five

<sup>3</sup><http://csrc.ucf.edu/data/UCF101.php>



types: Human-Object Interaction, Body-Motion Only, Human-Human Inter-  
 240 action, Playing Musical Instruments, and Sports. We perform evaluation ac-  
 cording to the three splits of training and test as described in [46] and present  
 exhaustive results on the dataset.

Although Faster R-CNN is the state-of-the-art object detection method,  
 there still exists lots of action frames that it cannot detect or only detect very  
 245 few persons. As person detection is a very important part of our method, small  
 number of sketches is not enough to find the most distinctive action sketches.  
 Considering the practical needs and computational cost, we set the max number  
 of top distinctive action sketches  $m = 20$  for each video. So the videos will be  
 removed from the original dataset if the number of person detected by Faster  
 250 R-CNN is less than the max number  $m$ . Table 1 gives the number of train/test  
 samples in the two public action datasets after removing some videos. In the  
 following parts, all experiments are conducted on the refined action datasets if  
 not specifically stated.

Table 1: The number of train/test samples in KTH and UCF101 human action datasets after removing the videos that Faster R-CNN detect very few persons.

	KTH	UCF101		
		split1	split2	split3
train	342	7708	7766	7805
test	181	3083	3025	2986
total	523	10791	10791	10791

## 6.2. Implementation details

255 For IDT based features, we choose the combined descriptors (HOG + HOF +  
 MBHx + MBHy) with default parameter settings and utilize the implementation  
 of Wang<sup>4</sup> to extract features from action videos. Regarding the feature encoding,  
 the Fisher vector which has shown empirically to give good results is adopted

<sup>4</sup>[http://lear.inrialpes.fr/~wang/improved\\_trajectories](http://lear.inrialpes.fr/~wang/improved_trajectories)

to represent the videos. For the training of Gaussian mixture model (GMM),  
 260 we randomly sample a subset of 256,000 features to learn GMMs and set the  
 mixture number  $K = 256$ . Finally, the power and  $L2$  normalization are applied  
 to normalize the obtained Fisher vectors of each descriptor type.

In the experiments, we take the linear support vector machine (SVM) which  
 is very efficient in dealing with large data sets as the action recognition classifier.  
 265 Specifically, we use the code of LIBSVM implemented by Chang et al. [47]  
 released on their website<sup>5</sup>. In the case of multi-class classification, we adopt  
 the one-vs-all approach and select the class with highest score. Besides, we  
 also fix  $C = 100$  as described in [27] which has shown good performance. For  
 different sketch pooling methods, we adopt 5-fold cross validation to find the  
 best parameters  $u$  and  $m$  on the training set.

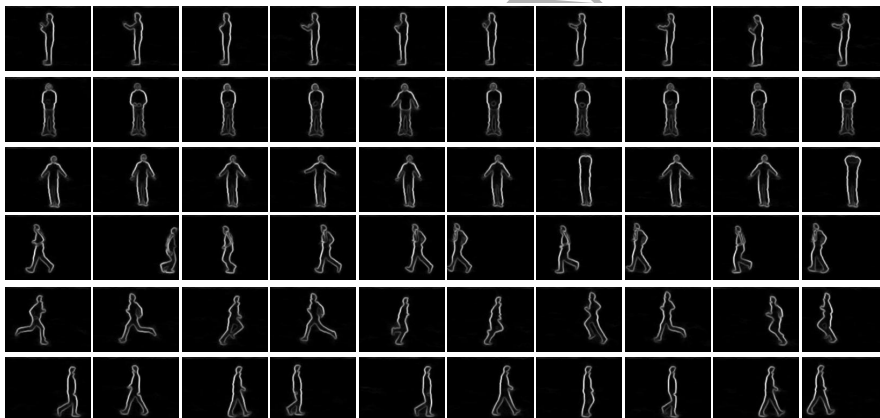


Figure 5: Examples of top 10 distinctive action sketches for each action class. From top  
 to bottom, they are boxing, hand clapping, hand waving, jogging, running and walking,  
 270 respectively.

### 6.3. Results of distinctive action sketch

We implement our method of distinctive action sketch mining on the six  
 action class of KTH dataset. Figure 5 shows examples of top 10 distinctive ac-  
 tion sketches for each action class. The experimental results demonstrate that

<sup>5</sup><http://www.csie.ntu.edu.tw/~cjlin/libsvm>

275 different action categories generally have diverse distinctive action sketches (ac-  
 tion patterns) and that our method performs well in capturing these sketches.  
 Through these vivid action sketches, the actions can be distinguished conve-  
 niently and efficiently. Another point emerged from Figure 5 is that sketches  
 of quick actions (walking, jogging, running) have more obvious variations than  
 280 the sluggish action (boxing, hand waving and hand clapping). Furthermore, we  
 can find that there exist many similar patterns in different categories of action  
 sketches. Actually, it is coincident with the intrinsic property in actions, for  
 that sequences between various actions usually share some analogous parts.

In the process of distinctive action sketch mining, a very important factor is  
 285 the cluster computation by k-means. To analyze the influence of cluster number  
 on final results, we carry out several experiments using different cluster numbers  
 on the KTH dataset. Figure 6 shows some results of distinctive action sketches  
 under different cluster numbers.

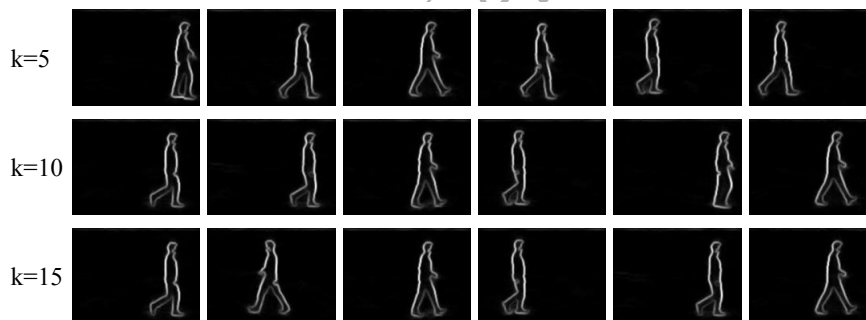


Figure 6: Results of distinctive action sketches under different cluster numbers. Take the case of walking, from top to bottom is top 6 distinctive action sketches under three kinds of cluster numbers ( $k = 5$ ,  $k = 10$ ,  $k = 15$ ).

290 It can be seen that a small number of cluster class yields weak distinctive  
 action sketches. The reason is that a relatively small cluster number will make  
 the results tend to on behalf of those sketches occurred frequently. Along with  
 the increasing of cluster number, we will get more similar distinctive sketches.  
 It means that there is no need to apply a too large number of clusters in pursuit  
 of the results' diversity. It not only cannot improve performance but also will

295 increase the computation cost. In the experiment, we set the cluster number  
 $k = 10$ .

#### 6.4. Comparison of different pooling methods

We test the performance of different pooling methods by taking the feature  
 vectors after pooling as the input of action classifier. Because the vectors do not  
 300 combine with IDT features, so we can evaluate the pure performance when using  
 them alone. Figure 7 gives a comparison of the four kinds of pooling methods  
 on KTH and UCF101 datasets. The parameter  $u$  in pooling process is varying  
 from 1 to 20 while top  $m = 20$  distinctive action sketches are discovered. We  
 can see that every pooling is the worst method for action recognition. That is  
 305 to say, it is very difficult to classify the videos represented by one action sketch  
 only. Other three pooling methods achieve good performance on the KTH and  
 UCF101 datasets. For a simple dataset like KTH, the variations of action and  
 backgrounds are very limited. It is possible to classify the action videos through  
 values with the maximum responses. But for the UCF101 dataset, the situation  
 310 is more complex. Average pooling takes the mean value of top distinctive action  
 sketches as the video representation. It can minimize the negative effect of action  
 variations. As the results shown, the max pooling and average pooling get the  
 highest accuracy on the two datasets respectively. It also demonstrates that the  
 top ranking action sketches are distinctive for action recognition.

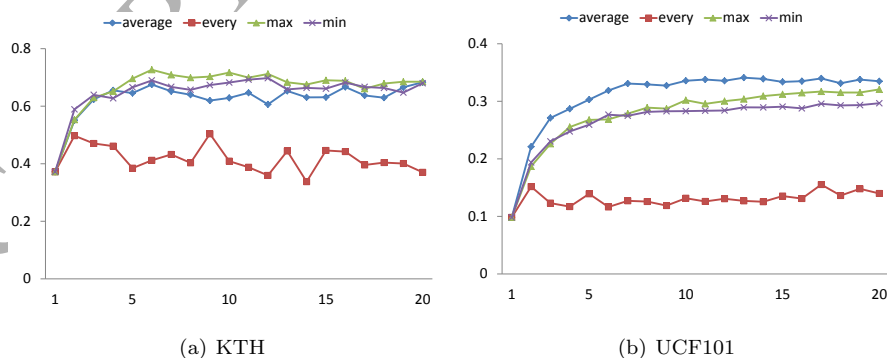


Figure 7: Comparison of different pooling methods with varying parameter  $u$  on the KTH and UCF101 datasets when  $m = 20$ .

315 Table 2 presents the performance of using different pooling methods alone  
 and gives the accuracies of combination with the Fisher vector of IDT features  
 on the KTH and UCF101 datasets. We can see that the accuracies of action  
 recognition are significantly improved after the representation-level fusion. It is  
 interesting to note that the every pooling get a surprising higher performance  
 320 than the other pooling methods on the KTH dataset. For the UCF101 dataset,  
 the combination of max pooling based action sketch and IDT features is the  
 best choice for action recognition.

Table 2: The classification accuracies of different pooling methods and combination with IDT features on the KTH and UCF101 datasets.

	average		every		max		min	
	sketch	comb.	sketch	comb.	sketch	comb.	sketch	comb.
KTH	68.32%	92.66%	53.88%	<b>94.58%</b>	72.67%	92.66%	70.44%	92.66%
UCF101	33.98%	83.53%	15.72%	83.59%	32.27%	<b>83.85%</b>	31.06%	83.09%

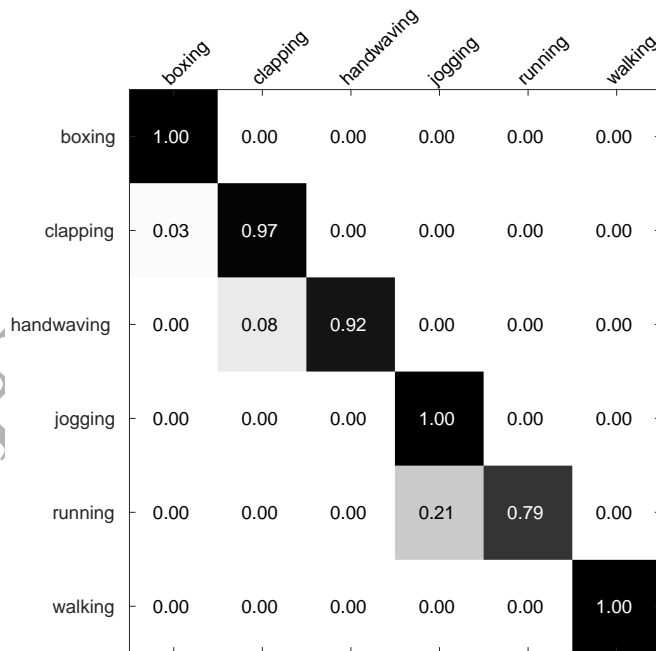


Figure 8: The confusion matrix of KTH dataset.

### 6.5. Results of action recognition

Table 3 shows the action classification accuracies of our method and IDT. As can be seen, the combination of sketch pooling and IDT leads to obvious performance gain. The average performance improvements are 3.03% and 1.42% for KTH and UCF101 datasets respectively. For the KTH dataset, the confusion matrix is shown in Figure 8. We can see that boxing, jogging and walking are perfectly recognized. For the three splits of UCF101 datasets, we also present the corresponding performance. As can be seen, the highest improvement 2.15% is achieved on the split2. Besides, Table 4 gives all the best parameters  $m$  and  $u$  selected for sketch pooling on the two datasets. The results of two human action datasets clearly demonstrate the effectiveness of our method.

Table 3: Comparison between our method and IDT on the KTH and UCF101 datasets.

	IDT	ours	improvement
KTH	91.55%	<b>94.58%</b>	+3.03%
UCF101	split1	<b>81.14%</b>	+1.19%
	split2	<b>82.99%</b>	+2.15%
	split3	<b>84.36%</b>	+0.90%
	average	82.43%	<b>83.85%</b>

Table 4: The best parameters  $m$  and  $u$  selected for sketch pooling.

KTH		UCF101		
		split1	split2	split3
$m$	11	10	18	20
$u$	6	4	3	20

We also show the comparison against several state-of-the-art methods of action recognition on UCF101 datasets. As a final experiment, we explore the performance of the proposed approach by a late fusion with enhanced motion vector (EMV) CNN [48] and Temporal Segment Networks (TSN) [49]. To get the final output of action video, we simply plus the SVM scores of our method with the predictions for each action class of EMV CNN and TSN. Furthermore,

340 the experiments above are conducted on the refined action datasets as we have mentioned in Section 6.1. For these videos that do not exist in the refined dataset, we directly take the predictions of EMV CNN and TSN as the final outputs. The results are shown in Table 5. Combined with TSN, our method obtains the state-of-the-art performance on UCF101 dataset (95.1%). Besides, 345 we achieve 3.2% performance improvement and get 89.6% compared to EMV CNN. It can be seen that our method can provide a significant complementary for the approach based on convolution network.

Table 5: Comparison of the performance with state-of-the-art methods on UCF101 dataset.

	Accuracy
Spatiotemporal CNN [50]	65.4%
Bimodal encoding [51]	84.2%
C3D(3 net) [52]	85.2%
Two-stream CNN [53]	88.0%
Factorized CNN [54]	88.1%
Two-stream+LSTM [55]	88.6%
TDD+IDT [56]	91.5%
Two-stream fusion+IDT [57]	93.5%
EMV CNN [48]	86.4%
TSN [49]	94.8%
Ours+EMV CNN	89.6%
Ours+TSN	<b>95.1%</b>

### 6.6. Running cost

350 The distinctive action sketch mining and pooling are the most important parts of our method. It takes roughly 4 hours for all videos in the refined UCF101 dataset, excluding the time of feature extraction. Benefits from the linear SVM, the training time on one split is only about 2.5 minutes while the test time can be negligible. The experiments are performed using Matlab 2014b

on a server configured with 24 Intel Xeon E5645 CPU and 64 G of RAM. The  
 355 detailed running time of different phases are listed in Table 6.

Table 6: Runtime (seconds) of different phases on UCF101 dataset.

Sketch mining	Sketch pooling	Training	Test
14167	342	150	1.5

## 7. Conclusion

Along with the booming development in computer vision, there is an increasing realization that sketch can be an essential element for many realistic applications. Unlike many previous works focusing the representation of action  
 360 sketch, we mainly explore the available transformation from action to sketch. Before implementation of the specific method, we analyze the characteristics that action sketch must meet and propose a reasonable framework of action sketching. Given the sketches generated under this framework, a distinctive ranking method is presented to mine the most representative sketches in action  
 365 videos. After that, we perform sketch pooling to obtain a new representation for action recognition. Experimental results demonstrate the effectiveness and excellent performance of our approach.

## Acknowledgements

The work was supported in part by the National Science Foundation of  
 370 China (61472103, 61772158, 61702136, and 61701273) and Australian Research Council (ARC) grant (DP150104645). We especially would like to thank the China Scholarship Council (CSC) for funding the first author to conduct the partially of this project at Australian National University.



**References**

- 375 [1] M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa, Sketch-based image retrieval: Benchmark and bag-of-features descriptors, *IEEE Transactions on Visualization and Computer Graphics* 17 (11) (2011) 1624–1636.
- [2] Y. Cao, C. Wang, L. Zhang, L. Zhang, Edgel index for large-scale sketch-based image search, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 761–768.
- 380 [3] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, S.-M. Hu, Sketch2photo: internet image montage, in: *ACM Transactions on Graphics*, Vol. 28, 2009, p. 124.
- [4] K. Xu, K. Chen, H. Fu, W.-L. Sun, S.-M. Hu, Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models, *ACM Transactions on Graphics* 32 (4) (2013) 123.
- [5] J. P. Collomosse, G. McNeill, Y. Qian, Storyboard sketches for content based video retrieval, in: *IEEE International Conference on Computer Vision*, 2009, pp. 245–252.
- 390 [6] R. Hu, S. James, T. Wang, J. Collomosse, Markov random fields for sketch based video retrieval, in: *ACM International Conference on Multimedia Retrieval*, 2013, pp. 279–286.
- [7] Z. Sun, C. Wang, L. Zhang, L. Zhang, Free hand-drawn sketch segmentation, in: *European Conference on Computer Vision*, Springer, 2012, pp. 626–639.
- 395 [8] Y. Zheng, H. Yao, S. Zhao, Y. Wang, Discovering discriminative patches for free-hand sketch analysis, *Multimedia Systems* (2016) 1–11.
- [9] M. Eitz, J. Hays, M. Alexa, How do humans sketch objects?, *ACM Transactions on Graphics* 31 (4) (2012) 44.

- 400 [10] Z. Sun, C. Wang, L. Zhang, L. Zhang, Query-adaptive shape topic mining for hand-drawn sketch recognition, in: ACM International Conference on Multimedia, 2012, pp. 519–528.
- [11] R. Poppe, A survey on vision-based human action recognition, *Image and Vision Computing* 28 (6) (2010) 976–990.
- 405 [12] S. Zhao, L. Chen, H. Yao, Y. Zhang, X. Sun, Strategy for dynamic 3d depth data matching towards robust action retrieval, *Neurocomputing* 151 (2015) 533–543.
- [13] A. Yilmaz, M. Shah, Actions sketch: A novel action representation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2005, pp. 984–989.
- 410 [14] P. Dollár, C. L. Zitnick, Structured forests for fast edge detection, in: IEEE International Conference on Computer Vision, 2013, pp. 1841–1848.
- [15] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
- 415 [16] Y. Zheng, H. Yao, X. Sun, X. Jiang, F. Porikli, Breaking video into pieces for action recognition, *Multimedia Tools and Applications* (2017) 1–18.
- [17] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, *Computer Vision and Image Understanding* 115 (2) (2011) 224–241.
- 420 [18] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, *Computer Vision and Image Understanding*.
- [19] Y. Ke, R. Sukthankar, M. Hebert, Spatio-temporal shape and flow correlation for action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
- 425

- [20] A. Klaser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference, 2008, pp. 275:1–10.
- [21] S. Zhang, H. Yao, X. Sun, K. Wang, J. Zhang, X. Lu, Y. Zhang, Action  
430 recognition based on overcomplete independent components analysis, *Information Sciences* 281 (2014) 635–647.
- [22] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: British Machine Vision Conference, 2009, pp. 124:1–11.
- 435 [23] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv:1304.5634.
- [24] W. Liu, H. Liu, D. Tao, Y. Wang, K. Lu, Multiview hessian regularized logistic regression for action recognition, *Signal Processing* 110 (2015) 101–107.
- [25] S. Zhao, H. Yao, Y. Gao, R. Ji, G. Ding, Continuous probability distribution  
440 prediction of image emotions via multitask shared sparse regression, *IEEE Transactions on Multimedia* 19 (3) (2017) 632–645.
- [26] I. Laptev, On space-time interest points, *International Journal of Computer Vision* 64 (2-3) (2005) 107–123.
- [27] H. Wang, C. Schmid, Action recognition with improved trajectories, in:  
445 *IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.
- [28] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
- 450 [29] W. Liu, D. Tao, Multiview hessian regularization for image annotation, *IEEE Transactions on Image Processing* 22 (7) (2013) 2676–2687.
- [30] L. Shao, L. Liu, M. Yu, Kernelized multiview projection for robust action recognition, *International Journal of Computer Vision* 118 (2) (2016) 115.

- [31] Y. Zheng, H. Yao, X. Sun, S. Zhao, Distinctive action sketch, in: IEEE International Conference on Image Processing, 2015, pp. 576–580.
- 455 [32] C. Hong, X. Chen, X. Wang, C. Tang, Hypergraph regularized autoencoder for image-based 3d human pose recovery, *Signal Processing* 124 (2016) 132–140.
- [33] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, *IEEE Transactions on Industrial Electronics* 62 (6) (2015) 3742–3751.
- 460 [34] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, *IEEE Transactions on Image Processing* 24 (12) (2015) 5659–5670.
- [35] H. Rahmani, A. Mahmood, D. Huynh, A. Mian, Histogram of oriented principal components for cross-view action recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (12) (2016) 2430–2443.
- 465 [36] R. Vemulapalli, F. Arraté, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 588–595.
- 470 [37] Z. Han, Z. Xu, S.-C. Zhu, Video primal sketch: A generic middle-level representation of video, in: IEEE International Conference on Computer Vision, 2011, pp. 1283–1290.
- [38] B. Yao, S.-C. Zhu, Learning deformable action templates from cluttered videos, in: IEEE International Conference on Computer Vision, 2009, pp. 1507–1514.
- 475 [39] C.-e. Guo, S.-C. Zhu, Y. N. Wu, Primal sketch: Integrating structure and texture, *Computer Vision and Image Understanding* 106 (1) (2007) 5–19.
- [40] Q. Fang, J. Sang, C. Xu, Giant: Geo-informative attributes for location recognition and exploration, in: ACM International Conference on Multimedia, 2013, pp. 13–22.
- 480

- [41] J. Sang, Q. Fang, C. Xu, Exploiting social-mobile information for location visualization, *ACM Transactions on Intelligent Systems and Technology* 8 (3) (2017) 39.
- [42] A. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, in: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.
- [43] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, T. M. Hospedales, Sketch-a-net that beats humans, in: *British Machine Vision Conference*, 2015, pp. 1–12.
- [44] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556*.
- [45] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: *IEEE International Conference on Pattern Recognition*, Vol. 3, 2004, pp. 32–36.
- [46] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, *arXiv preprint arXiv:1212.0402*.
- [47] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* 2 (3) (2011) 27.
- [48] B. Zhang, L. Wang, Z. Wang, Y. Qiao, H. Wang, Real-time action recognition with enhanced motion vector cnns, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2718–2726.
- [49] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: *European Conference on Computer Vision*, 2016, pp. 20–36.
- [50] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: *IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

- [51] J. Wu, Y. Zhang, W. Lin, Towards good practices for action video encoding, in: IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 2577–2584.
- [52] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
- [53] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014, pp. 568–576.
- [54] L. Sun, K. Jia, D.-Y. Yeung, B. E. Shi, Human action recognition using factorized spatio-temporal convolutional networks, in: IEEE International Conference on Computer Vision, 2015, pp. 4597–4605.
- [55] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.
- [56] L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4305–4314.
- [57] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.