Hand Detection and Tracking in Videos for Fine-grained Action Recognition

Nga H. Do and Keiji Yanai

Department of Informatics, The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan

Abstract. In this paper, we develop an effective method of detecting and tracking hands in uncontrolled videos based on multiple cues including hand shape, skin color, upper body position and flow information. We apply our hand detection results to perform fine-grained human action recognition. We demonstrate that motion features extracted from hand areas can help classify actions even when they look familiar and they are associated with visually similar objects. We validate our method of detecting and tracking hands on VideoPose2.0 dataset and apply our method of classifying actions to the playing-instrument group of UCF-101 dataset. Experimental results show the effectiveness of our approach.

1 Introduction

In recent years, using low-level features such as optical flows and spatio-temporal features to represent actions has become the most popular framework for action recognition. Several researches have employed high level features like human poses and human-object interactions. However, most of these high-level feature based approaches work only on still images and do not take advantage of motion characteristics of actions [1–4]. Among a couple of works which handle videos, Prest *et al.*'s work [5] learns human actions by using interactions between persons and objects. They proposed to localize in space and track over time both objects and persons, and represent actions as the trajectories of objects with respect to persons. Since their approach relies on object detection, they have to learn object detectors of all related objects. This process requires costly annotations. Moreover, they do not consider the case when the objects are visually similar. For instance, since violin and cello share the same visual characteristics, their detectors are supposed to fail to distinguish them. Consequently, it may be easy for Prest *et al.*'s method [5] to confuse "play violin" action and "play cello" action. In this paper, we propose to represent actions involved with objects solely based on how people perform them with the objects using their hands. According to our method, disparate actions associated with different but visually similar objects can be classified (see Fig.1 for the illustration). We show that hand related motion features are discriminative and representative enough for human actions.

In fact, in many cases of human actions, especially those that are involved with objects, people tend to move only their arms/hands to operate the actions.

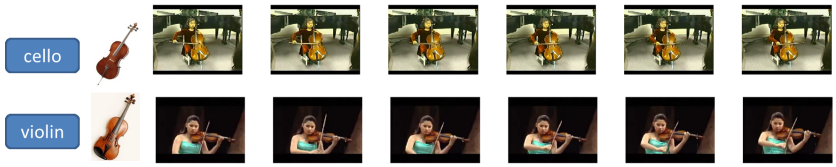

Fig. 1. An example which shows that actions with objects involved may not be recognized by object detection. The video shots are from UCF-101 dataset. We can see that cello and violin look very similar, since they are in the same class of musical instruments (string instruments). Therefore, it is not an easy task to distinguish the actions related with them ("play" in this case) using the two instrument detectors. However, while playing them, people put their arms/hands in different positions and move them in different directions. Consequently, exploiting motion features of arms/hands can be expected to be able to help classify "play cello" and "play violin".

In this paper, we focus on human actions which require hand movements during the time the actions are operated. Thus, we propose to take all possible arm/hand motions into consideration to represent the actions. Since motions of hands also contain those of arms, and in some cases, not the entire arm but only hands move, in order to handle as many cases as possible, here we focus only on movements of hands.

The detection of hands has been known as a tremendously challenging task since hands are the most flexible human body parts compared to others. Their appearance can change unpredictably since they can be closed or open, and the fingers can have various articulations. Moreover, in videos, they are naturally the fastest moving body parts. This means that unlike in still images where their shape can be recognized quite clearly, in videos sometimes they are very hard to detect due to motion blurs caused by their movements. In this paper, we propose to exploit multiple cues including hand shape, skin color, upper body position and flow information to detect hands in videos. Our objective is to obtain 2D+t sequences of bounding boxes which tightly bound hands in the videos. We demonstrate that using motion features extracted only from hand regions can achieve comparable performance to using motion features extracted from the whole frame. That means hand motion features are the most informative representation of human actions involved with hand movements. Moreover, we further enhance action recognition precision by exploiting displacement features of hands which belong to the most reliable hand tracks. To the best of our knowledge, we are the first to classify human actions using only hand related motion features.

Our contributions can be summarized as follows: (i) we develop an efficient hand detector for hand detection in videos based on multiple cues; (ii) we propose to recognize actions by exploiting the information of how people operate the actions with their hands; (iii) we propose a discriminative hand displacement feature which improves action recognition; (iv) we confirmed the efficiency of our proposed method on uncontrolled videos. To validate the effectiveness

of our hand detector, we use VideoPose2 dataset [1]. This dataset was originally developed for the challenging task of upper body estimation. To test the efficiency of our method on action recognition, we conduct experiments on playing-instrument group of UCF-101 [6] dataset. UCF-101 is one of the most challenging action dataset up to date with the large variations in human pose, object appearance, viewpoint, background and illumination conditions. Experimental results show the effectiveness of our approach.

The rest of this paper is organized as follows. Section 2 introduces some more related work. Section 3 describes our proposed method of hand detection and tracking. Section 4 explains how we apply the detecting and tracking results to action recognition. Experiments and discussions about their results are presented in Section 5. Finally Section 6 gives conclusions.

2 Related Work

Here we introduce some related work on detection and tracking of hands and some on recognition of actions involved with objects.

Hand detection and tracking: Hand detection is a topic which has a quite long history and a wide range of applications such as Human Computer Interaction, Sign Language translators, human pose recognition and surveillance. In the early stage of development, hand detection technique required markers or colored gloves to make the recognition easier. Second generation methods used low-level features such as color (skin based detection) [7, 8] or shape [9]. Most recent works on hand detection in videos are performed in 3D [10–13]. They employ depth information provided by depth cameras. As one in a few recent 2D hand detectors for videos, the hand detector proposed by Sapp *et al.* [14] exploits flow field. They propose to extract motion discontinuities by computing the gradient magnitude of the flow field, and learn a linear filter via SVM using this motion discontinuity magnitude cue specific to hands. Hands are detected as regions with the max response from the detector at each frame location over a discrete set of hand orientations. In their work, the results of hand detection are only used as additional cues for limb localization since their final purpose is not hand detection but upper body pose estimation.

Most of hand tracking methods assume that hands are the most moving objects in an image frame. In [15], Yuan *et al.* proposed to use a temporal filter to select the most likely trajectory of hand locations among multiple candidates obtained by "block flow" matching. In [16], Baltzakis *et al.* proposed a skin color based tracker which allows the utilization of additional information cues such as image background model, expected spatial location, velocity and shape of the detected and tracked segments. The benefit of their trackers is that they can track hands in real time. However, their trackers only work under constrained environments where the background is unchanged, so that simply substracting background can bring them enough cues to infer the most moving objects which refer to hands. In this paper, we detect and track hand in realistic videos where

[1] http://vision.grasp.upenn.edu/cgi-bin/index.php?n=VideoLearning.VideoPose2

maybe there are multiple moving objects and there also exists camera motion that can cause noise.

Recognition of actions related to objects: Approaches for recognizing object related actions have been developed for both static images and videos. Unlike works on static images [1, 3, 4], works on videos [17–19] generally take motion characteristics into account. Filipovych *et al.* [17, 18] modeled human-object interactions based on the trajectories and appearance of spatio-temporal interest points. Their approach was applied only to controlled videos taken from viewpoint of the actor by a static camera against a uniform background.

One of the most related work to ours is Gupta *et al.*'s work [19]. They employed hand trajectories to model the objects and the human-object motions for classifying interactions between humans and objects. In their work, the motion can be simply extracted based on background subtraction since they worked only on videos with constrained environment (static and fixed background). On the contrary, our approach tackles the problem in uncontrolled videos. Moreover, while Gupta *et al.*'s work requires annotation efforts for building training data, including the annotation for the locations of the objects in all video frames, our method does not require such time consuming efforts.

3 Hand Detection and Tracking

3.1 Hand detection

Here we aim to automatically estimate the hand locations using flow information and two trained detectors: a upper body detector and a static hand detector. As for flow estimation, we use DeepFlow proposed by[20]; for upper body detection, we employ Cavin's upper body detector [2]; for detection of static hands, we apply a state-of-the-art hand detector in still images proposed by Mittal *et al.*[21]. We improve their hand detector, originally developed only for hand detection in still images, to become a hand detector in videos by exploiting motion information and introducing upper body pose based spatial constraints.

Method of the baseline We, first, briefly describe the method of static hand detection in [21], which is used as the baseline of our hand detector. According to [21], hand hypotheses are first proposed by three dependent detectors: a sliding window hand shape detector, a context based detector, and a skin based detector. Then, the proposals are scored by all three detectors and a trained model for scores is used to verify them. The hand shape detector was trained using Felzenszwalb *et al.* [22]'s part based deformable model with HOG features. The contexts here refer to the cues captured around the hands, especially the wrists. In order to learn the contexts, another part based deformable model [22] was trained from the hand bounding boxes which were extended to cover the wrists. The skin detector, first, builds a skin mask based on the skin color of

[2] http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector/

face(s) detected by OpenCV face detector. It then detects skin regions by fitting lines using Hough transform and finding the medial axis of the blob-shaped regions. The hands are hypothesized at the ends of the lines.

The hand bounding boxes proposed by above three detectors are scored and combined as follows:

Hand detector score: the score obtained directly from hand detector.

$$\alpha_1 = \beta_{HD}(b) \tag{1}$$

where β_{HD} is the scoring function of the hand detector [22].

Context detector score: the score obtained by max-pooling over all bounding boxes which overlap with given hand boxes. The overlap threshold is set as 0.5.

$$\alpha_2 = \max_{b_h \in B_h} (\beta_{CD}(b_h)) \tag{2}$$

B_h refers to the set of context bounding boxes overlapping with the hand bounding box b_h. β_{CD} is the scoring function of the context detector [22].

Skin detector score: the score calculated by the fraction of pixels belong to skin regions in a given bounding box and denoted as α_3.

The three scores are combined into a single feature vector $(\alpha_1, \alpha_2, \alpha_3)$. This vector is then classified by a trained linear SVM classifier [23]. Finally, bounding boxes are suppressed depending on their overlap with other highly scored boxes using super pixel based non-maximum supperession. The superpixels are obtained by Arbelaez *et al.*'s method of image segmentation [24].

Mittal *et al.* trained their detector by using the data which was collected by themselves from various public image datasets including PASCAL VOC 2007 [3], PASCAL VOC2010 [4], Poselet [25], Buffy stickman [5], INRIA pedestrian [26] and Skin dataset [27], with 2861 hand instances for training and 660 hand instances for test in total. According to their experimental results, 48.2% of the test instances were correctly detected.

Proposed method Even though Mittal *et al.*'s hand detector achieved good performance, it needs two conditions about the data to work well: first, image resolution should be high and second, face should be easy to detect. Hands in images with good resolution commonly have clear shape, so that shape detector can be effectively employed. Moreover, most of faces in their test data can be seen from front view, so that face skin based hand detection is possible. However, here we have to deal with more complex and totally unconstrained data. In our data, many videos have low resolution or are taken under bad light condition, and faces are sometimes hard to be recognized. In such cases, we cannot find any of shape and/or color cues to detect hands. Thus, instead of employing Mittal *et al.*'s detector as it is, we propose to make it possible to work in such videos

[3] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/
[4] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/
[5] http://www.robots.ox.ac.uk/ vgg/data/stickmen/

by introducing upper body based spatial constraints and motion information. The pose and position of detected upper body are used for two purposes: to estimate face region and to refine final detection results. On the other hand, flow information is exploited in two directions: to select upper body and to rescore hand hypotheses.

Our proposed method of hand detection is a three-step method which can be summarized as follows: (1) Detecting upper body by employing upper body detector and motion information, (2) Finding hand hypotheses based on multiple static cues, (3) Inferring the best hand hypotheses by exploiting motion cue and upper body based spatial constraints. Refer to Fig.2 for the illustration of our proposed method of hand detection in videos.

For a given frame, at the first step, we apply Calvin *et al.*'s upper body detector and flow information to detect the most dominant upper body. This upper body detector has been demonstrated as a powerful human pose estimator and applied by many approaches recently. One of this detector's benefits is that it can estimate rather precisely the head position even when the face is hard to be detected. This detector returns several results, each of them contains position of head, torso and two limbs, and scores for each result. However, the problem is that not all results returned by this detector are perfect, and the good ones are sometimes not highly scored. Moreover, we found that even when the faces and the torsos are quite precisely localized, it is not going that well for the limbs.

Assume that there exists at least one good prediction among the results, we infer it by introducing motion information and spatial constraints. We postulate the two following holistic hypotheses: (i) hypothesis about hands: hands are the most moving body parts in a upper body, and generally looked not big compared to the upper body from common views; (ii) hypothesis about the main actor: the main actor is generally in motion and captured in the easiest way to recognize. That means his or her upper body is likely in the middle of the image frame, and/or bigger than the others. Based on the first assumption, "good" upper body should cover moving regions, and these regions are supposed to include hands. For each detected upper body, we first segment it to regions with different movements by the gradient magnitude of the flow field. Regions that are smaller than upper body area multiplied by predefined area threshold $thresh_a$ then become motion based hand hypotheses of that upper body. Score of a upper body is redefined as the ratio between areas of hand hypotheses which lie inside and outside that upperbody and finally normalized by area of that upper body. In the case that there are no significant movements (no moving regions with average flow magnitude being larger than flow threshold $thresh_f$), "good" upper body is simply selected based on the second assumption: the more centered and the bigger, the higher probility to be selected. In our experiments, $thresh_a$ is fixed as 0.5 and $thresh_f$ is fixed as 1.

The second step is hand detection based on multiple static cues using Mittal *et al.*'s detector. The face of chosen upper body is used to detect skin regions. At the final step, detection results of the previous step are rescored by introducing following flow score and upper body score.

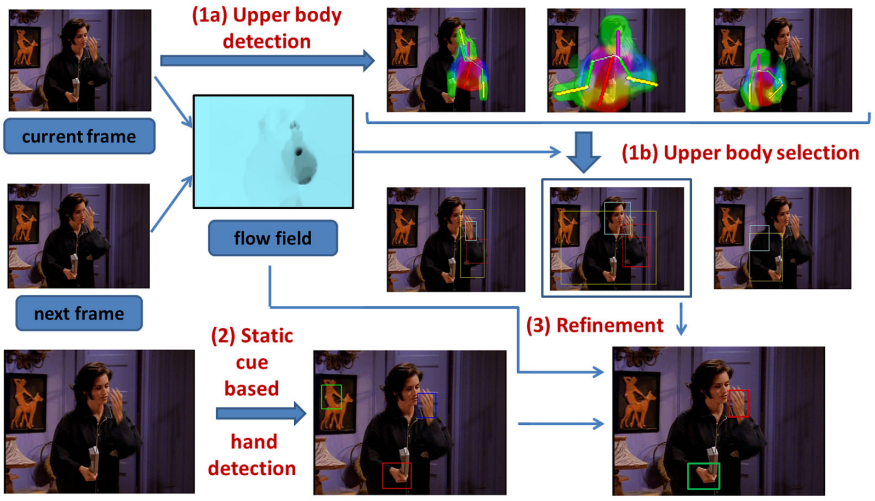

Fig. 2. Illustration for our proposed method of detecting hands. (1) For a given frame, we first apply Calvin *et al.*'s upper body pose estimator to obtain proposals of upper body pose. (1a) Each proposal consists of sticks with different colors: pink, red, green and yellow which respectively refer to position of head, torso, upper arm and lower arm. To infer the best detection, we employ the motion of flow field. We segment the frame by magnitude of flow field to obtain regions with disparate movements. For each detected upper body, its score is redefined as the ratio of regions which are supposed to be hands. These regions should, first, be in motion, and second, be not too large or too small compared to the upper body. We show these regions by the red bounding boxes. The detection at the right side contains no motion, thus it is not supposed to be the good one. (1b) The middle detection is considered to be better than the left one since it contains more motion based hand hypotheses. Upper bodies and faces are marked by yellow and light blue bounding boxes respectively. (2) The face of the selected upper body (the middle one), along with hand shape and context, are then used as static cues for hand detection following Mittal *et al.*. The three best results obtained by hand detection using static cues are shown in the middle image of the last row. The best is represented by the red bounding box, the second best is green and the third best is blue. As we can see, the second best is a failed detection even though it has hand shape. (3) We refine the detection results by considering motion information and upper body position and obtain the final results as shown in the right image of the last row.

Flow score: calculated as the average of gradient magnitude of flow of pixels in detection result. This score is normalized to have value between 0 and 1. We denote it as α_4.

Upper body score: determined by using spatial constraints based on position and area of upper body. It is calculated as percentage of area within the detected hand which overlaps with the upperbody. We also give penalties for detected hands that are too big compared to the upper body. For such detections, their upper body scores are fixed as -1. We denote upper body score as α_5.

The final score of a given bounding box is defined as follows:

$$\text{Mittal}'\text{s detector score} + w_f * \alpha_4 + w_u * \alpha_5 \tag{3}$$

w_f and w_u are weights for flow score and upper body score respectively and determined by experiments. We tried all values from 0.1 to 0.9. Based on our experiments, $w_f = 0.7$, $w_u = 0.2$ obtained the best performance.

3.2 Hand Tracking

In order to reduce the computational cost, we process hand detection for only one frame in every k frames. Thus, we need to track obtained detections and automatically link detecting and tracking results over time. We also want to compensate for missing detections as well as search for the most reliable hand tracks.

We track h highest scored bounding boxes of every detection through L frames forwardingly. Since we obtain one detection in every k frames, we need to consider $L * h/k$ bounding boxes. We capture the persistence of hands over time with simple flow based tracking. We take the average flow of a bounding box to propagate it from a frame to the next. A reliable bounding box should overlap with many others during its propagation. A track of a detected bounding box will be employed if the bounding box overlaps more than 50% with any of h bounding boxes of at least n frames which have hand detection processed among L frames. In our experiments, $h = 2, L = 15, k = 3, n = 2$. Some example results of our method of hand detection and tracking are shown in Fig.3. As shown in Fig.3, we are able to not only compensate missing or undone detections but also remove false detections.

4 Application on Action Recognition

Here we describe our approach of classifying action videos of a given action dataset employing the results of hand detection and tracking obtained by our above method.

4.1 Overview of our approach

We first apply our proposed hand detector on each video in the dataset. To reduce computational cost, we do not perform hand detection for all frames but

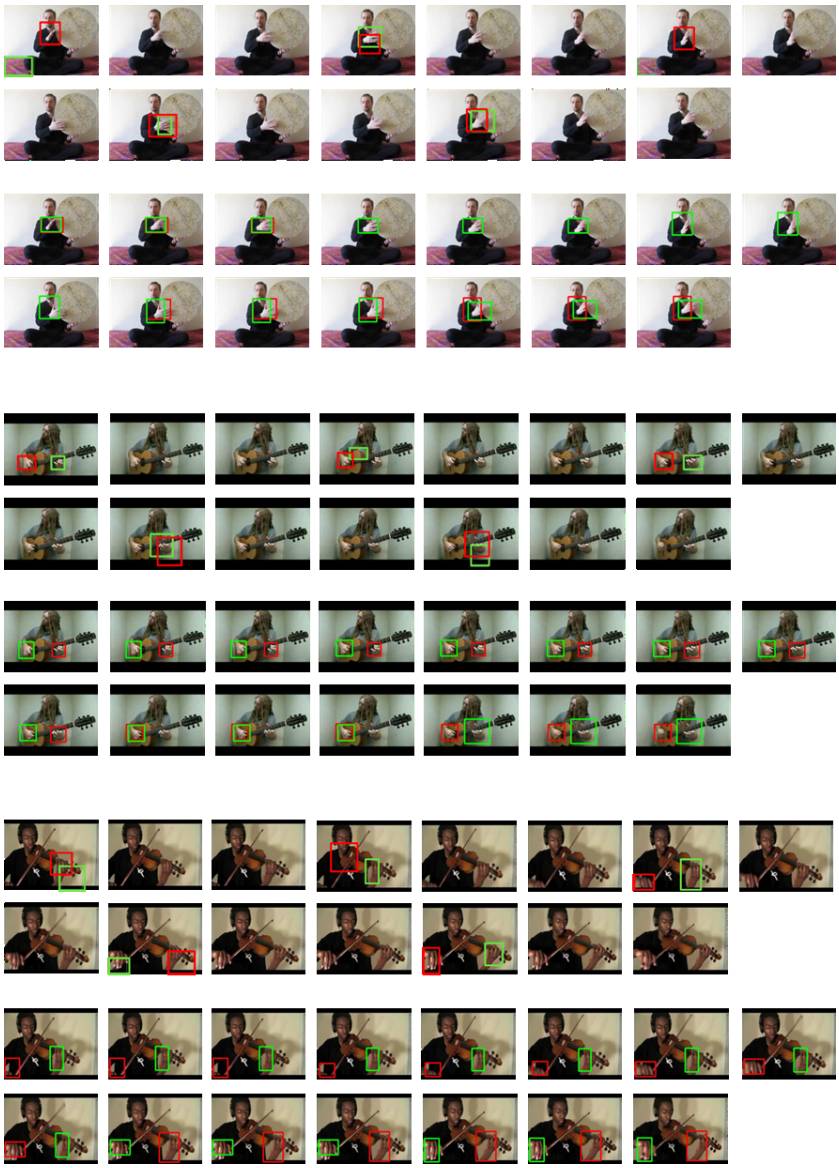

Fig. 3. Example results of our method of hand detection and tracking on the group of playing instruments in UCF-101 dataset. From the top, first the detection results, then the tracking results of "playing daf", "playing guitar" and "playing violin" are respectively shown. Among 15 consecutive frames, there are 5 frames which have hand detection processed. Only 2 top scored bounding boxes are shown. We track hands and keep hand tracks which overlap with at least 2 detection results. As the results, we can eliminate some failed detections and missing detections as well as obtain quite good hand bounding box sequences.

for only one in every k frames. In our experiments, k is set to 3. Next, based on the detection results, we track all highly scored bounding boxes through L frames to obtain connected and more reliable hand regions. In our experiments, L is set to 15. We then apply Wang *et al.* [28]'s method to extract dense trajectory aligned motion features and our hand displacement features from the detecting and tracking results. We conduct a Fisher vector for each type of extracted features. To combine different features, we concatenate their Fisher vectors. We train a multiclass linear SVM to classify the videos. By focusing only on regions which are expected to be the most related to the actions in stead of considering the whole frame, we can improve action recognigtion precision.

In this paper, we apply Fisher encoding methodology as described in [28]. The descriptor dimensionality is reduced by half using Principal Component Analysis (PCA). A subset of 256,000 features are randomly sampled from the training set to estimate the GMM and the number of Gaussians is set to 256. Each video is represented by a 2DK dimensional Fisher vector for each feature type, where D is the feature dimension after performing PCA. The following subsection explains in detail about the features.

4.2 Feature extraction

We extract features based on 2D+t sequences of hand bounding boxes obtained by our proposed method of detecting and tracking hands. We apply Wang *et al.*'s method [28] to extract dense trajectories and their aligned motion features: HOF (Histograms of Optical Flow) and MBH (Motion Boundary Histograms) from all detecting and tracking results. Their method recently became the state of the arts for action recognition. According to their method, dense trajectories are obtained by tracking sampled points using optical flow fields for multiple spatial scales. HOF and MBH descriptors are computed within space-time volumes around the trajectories. HOF directly quantizes the orientation of flow vectors. MBH splits the optical flow into horizontal and vertical components, and quantizes the derivatives of each component.

In this paper, we extract dense trajectories for only points which lie inside detected and tracked bounding boxes. If a frame has hand detection processed, its h highest scored detections will be used, otherwise, tracking results will be employed. We demonstrate that hand movements are discriminative and representative enough for actions operated by hands.

Beside dense trajectories and their aligned motion features, we extract hand track feature which describe the shape of hand trajectory by using average flow magnitude of hand regions in complete hand tracks. Given a trajectory of length L, its shape is described by a sequence $S = (\Delta P_t, ..., \Delta P_{t+L-1})$ of displacement vectors $\Delta P_t = (P_{t+1} - P_t)$. Here $P_t = (x_t, y_t)$ indicates the location of point P at frame t. The vector is normalized by the sum of the magnitudes of the displacement vectors:

$$S' = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||} \tag{4}$$

This vector is refered to descriptor for trajectory of point P. For our hand track features, only the center points of consecutive hand bounding boxes are taken into account. As the result, we obtain one descriptor for each hand track. While dense trajectories are extracted from all detection and tracking results, our descriptors are obtained from only reliable hand tracks. Thus even though they seem to be less informative than dense trajectories, they are expected to be a useful representation for actions as well.

5 Experiments

We conducted experiments to validate the efficiency of first, our method of detecting and tracking hands and second, our method of classifying actions based on our results of hand detection and tracking. Experiment results show the effectiveness of our approach.

5.1 Experiments on hand detection

Here we want to show how our proposed utilization of static cues and motion information can improve hand detection in videos. We compare detection performance between our detector, Mittal *et al.*'s detector [21] which uses only static cues and Sapp *et al.*'s detector [14] which employs only motion information.

We validated our proposed method of hand detection on VideoPose2.0 dataset. The dataset consists of 14 video shots collected from movie source. It was originally developed only for the task of upper body estimation. Therefore, the exact locations of hands are not provided. We had to annotate hands in every frame by ourselves. There are 2453 frames and 3814 hands in total.

In these experiments, we detected hands in every frame. The performance is evaluated using average precision following Mittal *et al.* [21]. A detection is considered true if its overlap score is more than 0.5. The overlap score of a detected bounding box B_d is defined as $O = \frac{area(B_g \bigcap B_d)}{area(B_g \bigcup B_d)}$, where B_g is the annotated grouth-truth bounding box. The results are summerized in Table 1 and some detection examples are shown in Fig.4.

First we validated the effectiveness of using faces of selected upper bodies instead of OpenCV face detector. As we can see in Table 1, the result was slightly improved. This is because VideoPose2.0 dataset has high resolution so that faces are usually big and clear enough for OpenCV detector to detect. The precision was significantly enhanced by introducing flow score. The first three rows of Fig.4 show the effectiveness of our detector over our baseline and flow based detector. While Mittal *et al.*'s detector sometimes failed to detect moving hands, mostly due to their unclear shape, our detector, by considering motion information, could detect them. This demonstrates that motion cue is extremely important for detecting hands in videos. However, employing only motion information can not robustly detect hands as Sapp *et al.*'s flow based detector could achieve only 18.6% precision. Their flow based detector only concentrates on detection regions moving similarly to trained hands. Our proposed method which utilizes

Fig. 4. Some examples of our detection results. We show two detections with best scores for each image frame. The best is shown in red, the second best is shown in green bounding box. The three upper rows of this figure show some detection examples in VideoPose2.0 dataset to compare the performance of the baseline, flow based detector and our detector (from the top, respectively). As we can see, our detector can detect more hands, especially hands blured by their movements. Especially, in the case that there are more than one character (the second image from the right), our detector tends to detect moving hands since they are expected to belong to the main character. On the other hand, using only static cues gives higher scores for static hands which may belong to the character in supporting role (the second example from the right). Using only motion cues (flow) concentrates on detecting moving regions (which sometimes belong to other body parts or background objects). The last row of this figure shows some detection results for the group of playing instruments in UCF-101 dataset.

static cues and motion information achieved the best results. By adding upper body based spatial constraints, the precision was further improved. Our method of hand detection improved the baseline approximately 5%.

5.2 Experiments on action recognition

Here we applied the results of hand detection and trackings to action classification. We aimed to classify actions based on how persons move hands to operate them. The actions should have hand movements involved throughout the time they are performed. However, there was too few public data which matches our purposes. We found only the group of playing instruments in UCF-101 dataset as suitable data for us to validate our method. UCF-101 is a very challenging action data set as its video shots are collected from Web source. The data set has 5 action groups, but only the group of playing instruments is suitable for the purpose of fine-grained action classification.

The group of playing instruments in UCF-101 dataset consists of 1428 video shots of actions of playing 10 types of musical instruments: cello, guitar, violin, daf, dhol, piano, tabla, sitar, flute and drum. The shots in each action category are grouped into 25 groups, where each group can consist of 4-7 shots of the action. The video shots from the same group may share some common features, such as similar background and similar viewpoint. We followed evaluation set up

Table 1. Results of hand detection. We conducted experiments on VideoPose2.0 video dataset and compared our hand detector with our baseline (Mittal *et al.*'s hand detector) and Sapp *et al.*'s flow based hand dtector. Our (+upper body) means using face of selected upper body for skin detector. Our (+flow) means adding flow score to refine detection results. Our (+flow+body) means using our full proposed method which employs both flow information and body position based constraints to improve the final results.

Method	Precision
Mittal *et al.* [21]	41.7%
Sapp *et al.* [14]*	18.6%
Our (+upper body)	42.6%
Our (+flow)	45.5%
Our (+flow+body)	46.3%

*Their flow based hand detector

as suggested in the ICCV2013 workshop on large-scale action recognition [6]. We adopted their provided three standard train/test splits to conduct experiments. In each split, clips from 7 of the 25 groups are used as test samples, and the rest for training. The result of each experiment reported here is calculated as the mean of average accuracies over the three provided test splits. We train multiclass linear SVMs [29] to perform action recognition.

For data from UCF-101 dataset, to reduce computational cost, we performed hand detection for only one in every three frames. To compensate the detections through the video as well as to find reliable hand tracks, we tracked hands as described in the subsection 3.3. Since UCF-101 is a large dataset without hand annotations, we could not validate the performance of our method of hand detection and tracking on this data in details. However, based on experimental results, we demonstrate that extracting features from regions specified to hands can achieve comparable performance to extracting from the whole frame. Our baseline in the experiments here refers to the method of extracting dense trajectories proposed by Wang *et al.* [28]. The results of our experiments are shown in Table 2.

As shown in Table 2, using only hand displacement features obtained 36% accuracy and using dense trajectories with their aligned motion features which were extracted from detected hand regions achieve comparable recognition performance to using original dense trajectories which were extracted from more regions. Even though precision rate of hand detection is not significant, imprecisely detected regions do not affect the final results that much since they are also informative (they are detected and employed by the baseline). The baseline, improved dense trajectory based method, extracts features only from foreground regions which move robustly. Instead of using all of those regions, in our method, we concentrate only on hand regions. The point is, despite of the fact that we use less information, we achieved comparable results to the baseline. That means our detection results are representative enough for the actions. Moreover, by combining multiple motion features considering hand positions, we could improve

[6] http://crcv.ucf.edu/ICCV13-Action-Workshop/

Table 2. Results of classification of actions in videos. DT means dense trajectories originally proposed in [28]. HDT means dense trajectories restricted to detected hand regions. HOF_{dt} and HOF_{hdt} refer to HOF features aligned with DT and HDT respectively (similarly with MBH). HT means our proposed hand track based displacement features. + means concatenating descriptors to a single descriptor before training and testing (early fusion).

Method	Precision
DT	66.7%
HOF_{dt}	83.8%
MBH_{dt}	86.6%
$MBH_{dt} + HOF_{dt} + DT$	87.3%
HDT	66.1%
HOF_{hdt}	81.4%
MBH_{hdt}	85.7%
$MBH_{hdt} + HOF_{hdt} + HDT$	86.2%
$\boldsymbol{HT}$	**36.0%**
$MBH_{dt} + HOF_{dt} + DT + HT$	87.6%
$\boldsymbol{MBH_{hdt} + HOF_{hdt} + HDT + HT}$	**88.5%**

the baseline. This result demonstrates that the proposed method can extract the features which have different characteristics from the conventional features. We also could prove that hand related motion features are particularly useful to recognize human actions.

6 Conclusions

In conclusions, we developed an effective hand detector in uncontrolled videos and obtained promissing results. Furthermore, we proposed to improve action recognition precision by additionally considering hand movements. Our experiment results showed that this consideration is effective. We try to deeply consider hand movements for the problem of improving action classification in uncontrolled videos. To the best of our knowledge, we are the first to do that. This is the largest contribution of this paper. Even if hand detection accuracy was only about 50%, employing the hand detection could help improve action recognition accuracy. This is a meaningful result even though the improvement is not remarkably significant. If hand detection accuracy is further enhanced, the benefit which action recognition gains from that enhancement can be expected to be larger.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 26011435.

References

1. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: Proc. of IEEE Computer Vision and Pattern Recognition. (2010) 17–24
2. Yao, B., Fei-Fei, L.: Discovering object functionality. In: Proc. of IEEE International Conference on Computer Vision. (2013) 2512–2519
3. Delaitre, V., Sivic, J., Laptev, I.: Learning person-object interactions for action recognition in still images. In: Advances in Neural Information Processing Systems. (2011)
4. Prest, A., Schmid, C., Ferrari, V.: Weakly supervised learning of interactions between humans and objects. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 601–614
5. Prest, A., Ferrari, V., Schmid, C.: Explicit modeling of human-object interactions in realistic videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 835–848
6. Khurram, S., Amir, R., Mubarak, S.: UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR **abs/1212.0402** (2012)
7. Binh, N.D., Shuichi, E., Ejima, T.: Real-time hand tracking and gesture recognition system. Proc. of International Conference on Graphics, Vision and Image Processing (2005) 19–21
8. Manresa, C., Varona, J., Mas, R., Perales, F.: Hand tracking and gesture recognition for human-computer interaction. Electronic letters on computer vision and image analysis **5** (2005) 96–104
9. Angelopoulou, A., Rodríguez, J.G., Psarrou, A.: Learning 2d hand shapes using the topology preservation model GNG. In: Proc. of European Conference on Computer Vision. (2006) 313–324
10. Ren, Z., Yuan, J., Zhang, Z.: Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In: Proc. of ACM International Conference on Multimedia. (2011) 1093–1096
11. Van den Bergh, M., Van Gool, L.: Combining RGB and ToF cameras for real-time 3d hand gesture interaction. In: IEEE Workshop on Applications of Computer Vision. (2011) 66–72
12. Cerlinca, T., Pentiuc, S.: Robust 3d hand detection for gestures recognition. In: Intelligent Distributed Computing V. Volume 382 of Studies in Computational Intelligence. (2012) 259–264
13. Oikonomidis, I., Lourakis, M.I., Argyros, A.: Evolutionary quasi-random search for hand articulations tracking. In: Proc. of IEEE Computer Vision and Pattern Recognition. (2014)
14. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. In: Proc. of IEEE Computer Vision and Pattern Recognition. (2011) 1281–1288
15. Yuan, Q., Sclaroff, S., Athitsos, V.: Automatic 2d hand tracking in video sequences. In: IEEE Workshops on Application of Computer Vision. Volume 1. (2005) 250–256
16. Baltzakis, H., Argyros, A., Lourakis, M., Trahanias, P.: Tracking of human hands and faces through probabilistic fusion of multiple visual cues. In: Computer Vision Systems. Volume 5008 of Lecture Notes in Computer Science. (2008) 33–42
17. Filipovych, R., Ribeiro, E.: Recognizing primitive interactions by exploring actor-object states. In: Proc. of IEEE Computer Vision and Pattern Recognition. (2008) 1–7

18. Filipovych, R., Ribeiro, E.: Robust sequence alignment for actor-object interaction recognition: Discovering actor-object states. Computer Vision and Image Understanding **115** (2011) 177–193
19. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009) 1775–1789
20. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: Large displacement optical flow with deep matching. In: Proc. of IEEE International Conference on Computer Vision. (2013)
21. Mittal, A., Zisserman, A., Torr, P.H.: Hand detection using multiple proposals. In: Proc. of British Machine Vision Conference. (2011) 1–11
22. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 1627–1645
23. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery **2** (1998) 121–167
24. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 898–916
25. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: Proc. of IEEE International Conference on Computer Vision. (2009)
26. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of IEEE Computer Vision and Pattern Recognition. Volume 1. (2005) 886–893
27. Karlinsky, L., Dinerstein, M., Harari, D., Ullman, S.: The chains model for detecting parts by their context. In: Proc. of IEEE Computer Vision and Pattern Recognition. (2010) 25–32
28. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proc. of IEEE International Conference on Computer Vision. (2013) 3551–3558
29. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. The Journal of Machine Learning Research **6** (2005) 1453–1484