VIEW AND SCALE INSENSITIVE ACTION REPRESENTATION AND RECOGNITION

Cao Yuanyuan, Huang Feiyue, Tao Linmi, Xu Guangyou

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ABSTRACT

In this paper a view and scale insensitive action representation VSI-Surf is proposed. Scale invariant shape descriptor R-transform is used to extract compact 1D feature from view insensitive posture representation “Envelop shape” which uses only two orthogonal cameras without accurate calibration. Considering action is a posture sequence, to integrate temporal information, 1D posture feature is then extended in time dimension. Then we get an action representation insensitive to viewpoint and scale, which is called VSI-Surf. Actions recognition is processed in a hierarchical framework, in which body actions and gestures are recognized in different level. Encouraging recognition results have been demonstrated on the multi-view IXMAS action dataset.

Index Terms—Viewpoint insensitive; action representation; action recognition

1. INTRODUCTION

Action recognition is an important research topic in computer vision. Over time, action recognition began to face challenges from a growing number of applications, such as intelligent surveillance, elderly monitoring, where the subjects are free to move in a real 3D environment. To recognize actions of subject is challenging due to the following reasons: (1) The Image features characterize the same action might be quite different in terms of both shape and scale due to the different viewpoints and distances; (2) Individual variations exist when different subjects performs the same action. To make the vision system applicable in real life, it is critical for action recognition method to handle the problems caused by view and scale variances.

3D models are widely used in action recognition to address the view-invariance issue[7,8]. However, inferring 3D poses from multiple calibrated cameras usually is computationally expensive due to the large number of parameters involved. Besides, recovered poses are often ambiguous under the perspective projection. The facts mentioned above make the 3D model based methods often not feasible for many applications.

On the other hand, the view centered description, namely 2D model based description possess the advantage of simple computation, but is of view and distance sensitive. To overcome this problem, the 2D model based methods are relied on models trained under wide range of view angles [2,3,6], which is similar to literature[7] in spirit. Nevatia[3] starts from rendering synthetic 2D poses from 36 viewpoints. Ogale[2] and Niu[6] also used atomic poses extracted from a set of multi-view multi-person video sequences, so as to be able to recognize actions under different viewpoints. However, one of the difficulties in using these methods is that synthesizing a series of 2D images and comparing them with the observed 2D images can incur a high computational cost; besides, the system needs to train and store a large number of data.

In the summary, the information from 2D images, even captured from multi-views, is not sufficient for the action recognition with complete view and distance invariant property. In this regard, for the 2D model based method, the essential issue is to find the representation which is not exactly view and distance invariant, but the variances are small enough in terms of discriminating different actions. In other words, the inner-class distances are much smaller than the inter-class distances between different actions. In this paper, 2D view and scale insensitive action representation based on only two cameras is suggested. First, a view insensitive 2D posture representation, named as “Envelop shape”[4], is computed. Training and storing a group of 2D templates from multiple views is not required anymore; yet the features insensitive to viewpoint and distinctive among different actions are preserved even using only two cameras. However, it is sensitive to silhouette scaling caused by distance changing between cameras and actors. This scaling problem is hardly discussed in the work above, neither. Even though, this is a practical problem we need to face. To solve this problem and integrate temporal information, VSI-Surf, a view and scale insensitive action representation is proposed. Compact low dimensional features are extracted on “Envelop shape” using R-transform and extended in time axis, which forms a surface. It is robust to noisy data and computationally efficient. VSI-Surf measures the average silhouette changing of the whole “Envelop shape”. However, for gestures like “wave”, only motion of arms contributes to action semantics; while for body actions like “sit down”, the

This work was supported in part by the National Natural Science Foundation of China under grants Nos. 60673189, 60873266 and 90820304.
body movement needs to be considered as a whole. Therefore, our recognition framework is designed to be a coarse-to-fine procedure. First, body actions are recognized and then gestures are discriminated. Support Vector Machine (SVM) is used to train and recognize actions.

The advantages of our method are: 1) only two orthogonal cameras without accurate calibration are needed; 2) The algorithm is computationally efficient; 3) 2D scale and view insensitive representation which is highly compressed simplifies action recognition by decreasing a large amount computation.

2. POSTURE REPRESENTATION

To represent an action, both spatial and temporal information are needed. The view insensitive posture representation “Envelop shape” is employed to represent special features in the assumption that view point only changes around vertical axis. To compute “Envelop shape”, two orthogonal cameras with the image plane parallel to vertical axis are needed. This assumption is suitable in most of action recognition situations. “Envelop Shape” has several advantages: (1) It is easy to obtain. Only silhouettes are required as input, which are easier to extract than meaningful feature detection, tracking or point correspondence; (2) It needs only two cameras which is feasible in applications; (3) The posture in each time is represented by a 1D vector regardless of the viewpoint, which saves a great of efforts on training and storing.

"Envelop shape" is computed using the images recorded from two cameras. In each horizontal section plane of “Envelop shape”, the radius is calculated using equation (1)

\[ r = \sqrt{x^2 + y^2} \]  

where \( x \) is body silhouette radius of one camera; \( y \) is body radius of the other camera. An example of “Envelop shape” for action “sitting down” is shown in Fig. 1.

![Fig.1 The first and the second rows are two synchronized video sequences showing action “sit down”; the third row shows “Envelop shape”.](image)

Experiments show the requirement for orthogonal cameras can be relaxed without much influences on the results, therefore accurate calibration of camera is not necessary. Let’s assume the angle of two camera optical axis is \( \pi - \alpha \), then the angle of two image plane is \( \alpha, r_0 \) denotes the minimum \( r \) in all rotation angles. The change of \( r/r_0 \) is shown in Fig.2. It is worth noting that when \( \alpha \) falls into the range of \([\pi/4, 3\pi/4]\), value of \( r/r_0 \) is between 1.15 and 1.414.

![Fig. 2 \( r/r_0 \) curve. The x axis represents \( \alpha \), the y axis represents \( r/r_0 \)](image)

3. ACTION REPRESENTATION: VSI-SURF

An action is a posture sequence along time axis. “Envelop shape” offers a posture representation in spatial dimension. Based on which \( R \) transform[5] is employed to integrate temporal information and generate an action representation VSI-Surf. \( R \) transform is a recently developed shape descriptor and has several advantages compared to competing descriptors. Firstly, it is computationally efficient and robust to noisy data. Therefore extremely accurate human silhouette is not necessary. Secondly, the \( R \) transform is invariant under translation and scaling which means it performs robust when the distance between subject and camera changes. Thirdly, it transforms a 2D shape to a compact 1D signal which is easy to implement.

The \( R \) transform converts a body silhouette to a 1D signal through the use of the two-dimensional Radon transform[5]. For an image \( I(x, y) \), the Radon transform, \( g(p, \theta) \), using polar coordinate \( (p, \theta) \), is defined as:

\[ g(p, \theta) = \sum_x \sum_y I(x, y) \delta(x \cos \theta + y \sin \theta - p) \]  

where \( \delta(.) \) is the Dirac delta function, \( \theta \in [0, \pi] \). However, Radon transform is sensitive to scaling, translation and rotation, the \( R \) transform extends Radon transform by calculating the sum of the squared Radon transform values for all of the lines of the same angle, \( \theta \), in an image:

\[ R(\theta) = \sum_p g^2(p, \theta) \]  

\( R \) transform has several favorable characteristics for the task of action recognition. Translation invariance allows human silhouettes appear in any position in the image frame without any affection on the results. Besides, when normalized, the \( R \) transform is scale-invariant. Scaling the silhouette image results in amplitude scaling of the \( R \) transform, so for our work, we use the normalized transform:

\[ \bar{R}(\theta) = \frac{R(\theta)}{\max_{\theta} R(\theta)} \]  

Through \( R \) transform, we get a compact representation with 180 dimensions instead of a 2D image frame, which still preserves important information to distinguish different actions. Fig.3 shows an example of the "envelop shape" and its \( R \) transform results. Considering the “Envelop shape” is vertically symmetrical, therefore:

\[ R(\theta) = R(\pi - \theta) \]  

So a feature vector with the length 90 is enough to describe the posture.
In previous work using R transform [1], action was regarded as a sequence of individual poses, and the recognition relied on pose-to-pose similarity measures. Since an action consists of spatial-temporal data, the temporal information plays a crucial role in recognizing action, which is ignored in a pose-to-pose approach. Therefore we integrate temporal characteristics by concatenating R curves along the time dimension, through which VSI-Surf is derived. The VSI-Surf for a sequence of silhouette image $I(x, y, t)$ is defined as:

$$ S(\theta, t) = \overline{R}_t(\theta) $$

where $\overline{R}_t(\theta)$ is the normalized R transform for frame $t$.

Some examples of VSI-Surf of different actions are shown in Fig.4. Fig.5 shows the VSI-Surfs of same action “kick” from 2 actors under 4 different viewpoints. As we can see from the figure, the appearance differences caused by individual and viewpoint are greatly eliminated.

4. VSI-SURF BASED ACTION RECOGNITION

VSI-Surf is a statistical measurement of the average silhouette changing of the whole body. In the ten actions to be recognized, some actions like “sit down”, “pick up”, “stand up”, and “kick” are whole body actions, which can be well described by computing VSI-Surf on “Envelop shape”. However, in some gesture actions, like “punch”, “scratch head”, “wave”, “cross arms”, and “check watch”, only the movements in upper part of the body should be measured to distinguish them. Based on this consideration, VSI-Surf on the whole body could heavily decrease the differences between gestures. Therefore, a hierarchical coarse-to-fine recognition strategy is proposed. In coarse phase, VSI-Surf of all actions are computed to recognize body actions, like “sit down”, “pick up”, “stand up”, and “kick”; all gestures are supposed to be recognized as “standing” in this phase. Then in the fine phase, actions in “standing” are further classified into different gestures by applying VSI-Surf on upper half of the body silhouettes. The proposed hierarchical recognition framework takes advantage of the semantics indicated by motion part of body and focuses on the most distinctive and contributive features, which improves the recognition results.

In this hierarchical framework, Support Vector Machine (SVM) with linear kernel is employed to learn and recognize actions. First all the surfaces are normalized to $180 \times T$ ($T=50$ in the experiment) matrix so that the action descriptor is robust to the duration of an action.

5. EXPERIMENTS

Our action recognition were conducted on the publicly available IXMAS dataset[9], which is also used by Weinland[7] and Yan[8] in a similar context. In the dataset actions are performed by 10 actors in arbitrarily position and orientation, each 3 times, and viewed by 5 calibrated cameras. Fig.6 shows five actors acting “kick” under five viewpoints (1) ~ (5). It is obvious that even the same action is performed in different ways by different actors.

It is worth noting that the viewpoints of recorded actions are actually unknown and the dataset contains actions in arbitrary views. We can see in Fig.6, two pairs of cameras (1)(2) and (3)(4) loosely meet the requirement of our “envelop shape” algorithm, which are called “camera pair”, though the image plane of camera (4) is not vertical. Besides, the actors performed all actions in three different orientations respectively. Thus we have 6 different view pairs which can be used to train and test the proposed approach. Moreover, the different scales of the silhouettes are suitable to test scale-invariance.

In our experiments, leave-one-out strategy is employed. Five of the view pairs are used for model learning; the remaining view pair is used for testing. This procedure is repeated by permuting the test view pair and the average recognition rate is computed. This experimental scheme is able to test view invariance attributes of the proposed action representation. Unlike Weinland[7], we report average results on all different views instead of the three best, which are shown in Fig.7. Though we used simple recognition method, for body actions like ‘stand up’, ‘sit down’, and ‘pick up’, the recognition results are consistently high. We believe it is because that pose changing in body level causes more intensive shape deformation during the process than gestures, which can be better described by VSI-Surf. Gestures like “wave”, “scratch head”, and “check watch”
are confused to each other, which is common in the results presented by other work. The reason is some of the detailed hand movement information has been discarded to achieve view insensitive representation.

Recognition results under 4 different view pairs. View1, 2 are under camera pair (1)(2) and View3,4 are under camera pair (3)(4).

The results are also compared with latest results reported in [8] which used 3D models, as shown in Fig.8. Our results are impressively better in body action like “pick up”, “kick”, “punch”, “stand up”, and “sit down” and comparable in other actions. The results are generally good considering that only simple 2D features are used instead of 3D models. Recognition results on combined multiple cameras are also reported as 71% for camera pair (1) and (2), 71% for camera pair (3) and (4) in [8]; For comparison, our recognition rates is 80% and 75.5% respectively.

Recogntion results of 4 different view pairs are 80%, 80%, 76%, and 75% respectively. We can see that the recognition performances are a little better under the camera pair (1) and (2). That’s because these two cameras better meets the camera configuration requirement of “Envelop shape” that two cameras should be orthogonal with image planes vertical than the other camera pair (3) and (4). However, the inspiring results also prove that the approach is not sensitive to the configuration of cameras; even it is derived from the assumption that

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present a view and scale insensitive action representation based on the video streams captured from a pair of cameras. A coarse-to-fine action recognition framework is proposed to recognize actions in body level and limb level hierarchically. The approach was evaluated on a public dataset of 10 actions and with different challenging scenarios. The inspiring results proved that VSI-Surf robustly captured the view and scale independent features of actions from arbitrary viewpoints and still computationally efficient. The proposed approach is suitable to be applied in real life applications like the elderly monitoring and assistant living systems, which is our future work.

7. REFERENCES