# Viewpoint Insensitive Action Recognition Using Envelop Shape

Feiyue huang, Guangyou Xu

Department of Computer Science and Technology,
Tsinghua University, 100084, Beijing, China
Hfy01@mails.tsinghua.edu.cn
xgy-dcs@mail.tsinghua.edu.cn

**Abstract.** Action recognition is a popular and important research topic in computer vision. However, it is challenging when facing viewpoint variance. So far, most researches in action recognition remain rooted in view-dependent representations. Some view invariance approaches have been proposed, but most of them suffer from some weaknesses, such as lack of abundant information for recognition, dependency on robust meaningful feature detection or point correspondence. To perform viewpoint and subject independent action recognition, we propose a representation named "Envelop Shape" which is viewpoint insensitive. "Envelop Shape" is easy to acquire from silhouettes using two orthogonal cameras. It makes full use of two cameras' silhouettes to dispel influence caused by human body's vertical rotation, which is often the primary viewpoint variance. With the help of "Envelop Shape", we obtained inspiring results on action recognition independent of subject and viewpoint. Results indicate that "Envelop Shape" representation contains enough discriminating features for action recognition.

**Keywords:** Viewpoint Insensitive, Envelop Shape, Action Recognition, HMM

## 1 Introduction

Human action recognition is an active area of research in computer vision. There have been several surveys which tried to summarize and classify previous existing approaches on this area [1], [2], [3], [4]. In this paper, we develop a general approach to recognize actions independent of viewpoint and subject, focusing on discovering viewpoint invariance for action recognition.

In our research work, we define human action recognition system as a system made up of three modules: Preprocessing, Pose Estimation and Recognition. Preprocessing module includes human detection and tracking, it extracts low level representation for pose estimation. Pose Estimation module is the process of identifying and representing how a human body and/or individual limbs are configured in a single frame. Recognition module uses results of pose estimation of frames to classify actions. Here we define posture as a kind of representation of human body in a single frame, for example, horizontal and vertical histograms of silhouette [5], vector of

distances from boundary pixels to the centroid [6]. In our opinion, posture representation is one of the most basic and key issues in action recognition system.

It is well known that a good representation for classification should have such measurement property whose values are similar for objects in the same category while very different for objects in the different categories. So this leads to the idea of seeking distinguishing features that are invariant to irrelevant transformations of the input [7]. In the case of recognition of human action, we argue that a good feature representation should be able to tolerate variations in viewpoint, human subject, background, illumination and so on. Among them, the most important invariance is viewpoint invariance. We can perform training and recognition according to given environments and specialized persons. But in order to perform natural human action recognition, we can't limit human body's movement and rotation at any time which inevitably leads to variable viewpoint.

It is a challenge to find a viewpoint invariant posture representation for action recognition. There have been some proposed approaches on viewpoint invariant action recognition. Campbell et al. proposed a complex 3D gesture recognition system based on stereo data [8]. Seitz and Dyer described an approach to detect cyclic motion that is affine invariant [9]. Cen Rao did a lot of research work on view invariant analysis of human activities [10], [11]. He used trajectory of hand centroid to describe an action performed by one hand. He discovered affine invariance of trajectory and his system can work automatically. Vasu Parameswaran also focused on approaches for view invariant human action recognition [12], [13]. He chose six joints of the body and calculated their 3D invariants of each posture. So each posture can be represented by a parametric surface in 3D invariance space. Daniel et al. introduced Motion History Volumes as a free viewpoint representation for action recognition, which needs multiple calibrated cameras [14].

Though there have been some research work on viewpoint invariant action recognition, there are still many problems to be solved. Most approaches depend on robust meaningful feature detection or point correspondence, which, as we know, are often hard to implement. And there is a tradeoff to be insensitive to viewpoint is that some useful information for discriminating different actions is often eliminated.

How to make representation insensitive to viewpoint while still keeping appropriate discriminating information for recognition appears to be the key issue. In this regard, we propose a posture representation named "Envelop Shape". Under the assumption of affine camera projection model, we prove it from both theory and experiments that such representation is viewpoint insensitive for action recognition. "Envelop Shape" is easy to acquire from low level features, which can be obtained from silhouettes of subjects by using two orthogonal cameras. It conveys more information compared to previous view invariant representation for action recognition. And it does not rely on any meaningful feature detection or point correspondence, which as we know is often difficult and sensitive to errors. With the help of our proposed representation, we develop our action recognition system. Experiment results show that our system has impressive discriminating ability for actions independent of subject and viewpoint.

The remainder of this paper is organized as follows. In section 2, we present our viewpoint insensitive representation. We propose the implementation of our action recognition system and give experiment results in section 3. We conclude in section 4.

## 2 View Invariant Posture Representation

In human action recognition, representation is a basic and key issue. A good system should have viewpoint invariance to perform natural action recognition. So a direct idea is to discover viewpoint invariant representation which means that measurements using this representation will keep almost the same even under different viewpoints.

### 2.1 Viewpoint in Action Recognition

Viewpoint transformation can be separated into two parts, translation and rotation. In action recognition, almost all representations have translation invariance, so we only consider rotation invariance. Figure 1 shows the coordinate in our system. In this coordinate, the Y-axis is vertical. There are three kinds of rotation terms used to describe the rotation in the coordinate: roll, pitch and yaw. Roll, pitch and yaw describe the rotation around the Z-axis (α), X-axis (β), and Y-axis (γ), respectively.

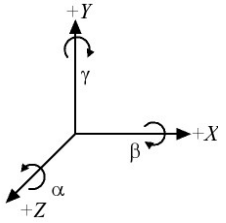


**Fig. 1.** Coordinate in our system.

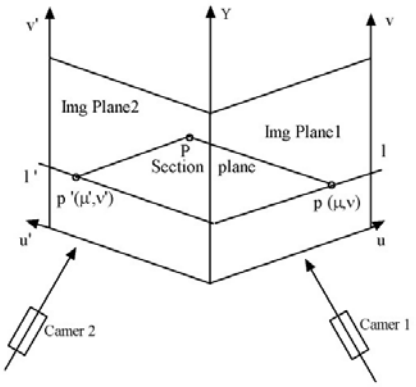


**Fig. 2.** Two cameras' configuration

It is quite often that actor makes some kind of action with his body yawing, when roaming in front of the fixed camera, for example the actor on the stage or teacher in front of the blackboard. In this case the yaw motion of the body causes variation of the viewpoint, but actions still make sense. In this regard, we classify human postures into a same category if only yaw rotation exists and we classify human postures in different categories if there exist the other two kinds of rotation terms, roll or pitch.

For example, when a human is standing compared with lying on the ground, the rotation term is roll or pitch, and we regard them as different postures. However, if a human only turns his body facing another direction, he is thought to be acting the same posture. With above discussion, we can conclude that we need only consider invariance on yaw rotation for most viewpoint invariant action recognition.

## 2.2 Envelop Shape Representation

In the practical situation of human action recognition, because the depth range of human body is usually small compared with the distance between human and the camera, the affine camera model can be used. To acquire viewpoint invariant representation for action recognition, a two cameras' configuration is proposed as Figure 2. The image planes of two cameras are both parallel to the vertical axis Y, and the optical axes are orthogonal. Let us consider a horizontal section plane of human body, projections of all points on this section plane into the image plane 1 are on the line l and projections of all points on this section into the image plane 2 are on the line l'. The line l are the epipolar line of point p' and the line l' are the epipolar line of point p. To discover the yaw rotation invariance, we need only to analyze the 2D horizontal section shape's projection on X-axis and Y-axis in different rotations.

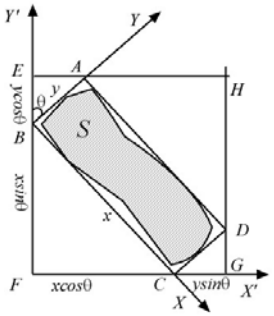


**Fig. 3.** 2D Shape projection on X-axis and Y-axis in different rotations

In Figure 3, let us suppose a 2D shape "S" whose projection segments in original coordinate XY are AB and BC, so it is in the rectangle ABCD. In another coordinate X'Y' which rotates at an angle θ, its projection segments will be in segment EF and FG. Let us define the original projection segment's length as x and y, the new projection segment's length as x' and y'. We can get the following relationships:

$$x' \le x\cos\theta + y\sin\theta \quad y' \le y\cos\theta + x\sin\theta \tag{1}$$

Let us define value "r" as equation 2, so we can get equation 3.

$$r = \sqrt{x^2 + y^2} \tag{2}$$

$$r' = \sqrt{x'^2 + y'^2} \le \sqrt{x^2 + y^2 + 2xy\sin 2\theta} \le \sqrt{x^2 + y^2 + 2xy} \le \sqrt{2}r \tag{3}$$

Let $r_0$ be the minimal value of "r" s among all rotations. Then at any rotation, the r value will meet the following expression:

$$r_0 \le r \le \sqrt{2} r_0 \tag{4}$$

This is a quite small value range compared to the unlimited range of ratios between x' and x or y' and y, which indicates that we find a view insensitive representation of human body. At each horizontal section plane, we can calculate an "r" value using equation (2). From a single frame of human posture, we get a vector of "r" value. Since this vector can envelop the human body silhouette inside, we call this representation of vector of "r" values as "Envelop Shape". Here we show some "Envelop Shape" images of synthesized human body model data at different viewpoints. Figure 4 shows two kinds of synthetic postures rotated on vertical axis Y at eight different angles, the first two rows are silhouettes of images in two cameras, and the third rows are "Envelop Shape" images. We can see that the Envelop Shapes does really change few facing viewpoint variance.

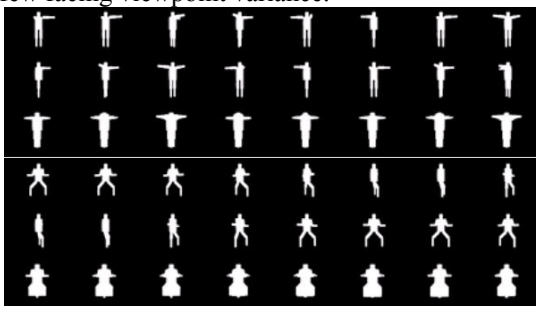

**Fig. 4.** Two kinds of postures at different viewpoints

Though we propose our cameras' configuration as Figure 2, that is, two cameras should be placed with image planes both parallel to the vertical axis Y and optical axes orthogonal. It does not need accurate calibration. As we know, accurate calibration is often complex. It is enough when the cameras are placed approximately meeting this need of cameras placement. That means we do not need to spend a lot of time in configuring the accurate placement. As we mentioned above, this kind of representation is just view insensitive, so approximate value can also work. We will show our experiments in section 3.The videos are collected with just rough placement of two cameras, while we will see that the result is still inspiring good.

Here is a brief description of our algorithm to generate Envelop Shape representation.

1. Extract silhouettes of human body from two cameras' video data.
2. Perform a scale normalization of the silhouettes.
3. Use expression (2) to calculate "r" value at each height of the silhouettes, the x and y are the corresponding width of the two normalized silhouettes at this height.

Envelop Shape representation has following advantages:

1. It keeps information on two dimensional degrees of freedom, vertical axis and horizontal plane. It has more information than some simple view invariance representation such as trajectory projection which is in fact one dimensional. Therefore, it has better discriminating ability, also it is view insensitive.

2. It is easy to obtain. Only silhouettes are required as input, which are easier to extract than meaningful feature detection, tracking or point correspondence.

## 3 Action Recognition Experiment Results

With the help of Envelop Shape representation, we deploy our arbitrary viewpoint action recognition system in a smart classroom. Figure 5 shows system flow diagram. We first use "Pfinder" algorithm to extract human silhouette [15]. With two cameras' silhouette video sequences as original input, we generate "envelop shape" vector of each frame. We then use principal component analysis (PCA) to perform dimensional reduction.

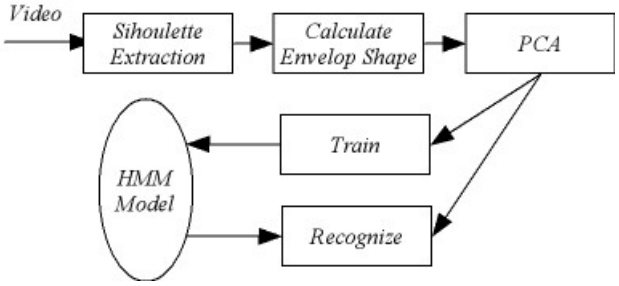


**Fig. 5.** Action Recognition System Flow Diagram

For each video, after preprocessing and posture representation steps are completed, time sequential feature vectors can be obtained. There are many algorithms for classifying time sequential feature vectors, such as Hidden Markov Model[16], Coupled Hidden Markov Model[17] or stochastic parsing[18] and so on. Here we use continuous Hidden Markov Model for action training and recognizing.

For action recognition experiment, we collected our own database of action video sequences. It contains seven different actors. Each actor performs nine natural actions which are "Point To", "Raise Hand", "Wave Hand", "Touch Head", "Communication", "Bow", "Pick Up", "Kick" and "Walk". Each action is performed by every actor three times repetitively at three arbitrary view point. Figure 6 and Figure 7 show examples of our experiment data. Each figure contains two groups of sampled action sequences. Each group contains five rows, the first two rows are images of two cameras, the next two rows are silhouettes extracted using "Pfinder" and the last row is normalized Envelop Shape vector images. (Each action sequence contains about 30 frames of images. Figure 6 and 7 only show partial sampled frames of each sequence.) As we can see, action sequences are collected at arbitrary view point, which means that our experiment is viewpoint independent.

Experiment parameters are as follows: the dimension of input vector (which is, output after PCA dimensional reduction of Envelop Shape vector) is 8, the number of states of each HMM Model is 5 and the number of mixture Gaussian Models for observation is 10. With our video database, we carry out subject dependent and subject independent experiments separately.

**Fig. 6.** Two groups of "point to" action sequences

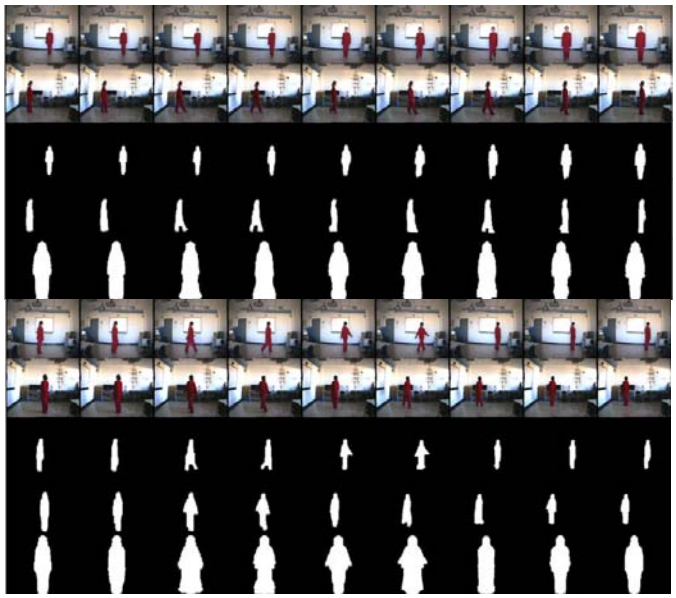

**Fig. 7.** Two groups of "walk" action sequences

In the case of subject independent action recognition, for each type of actions, we train an entire HMM models for all actors in train set and use the entire HMM models for recognition. For each action, we use first five actors' video sequences as training

sets and the last two actors' sequences as test sets. That is to say, for each action, there are 45 sequences as train sets and 18 sequences as test sets. Table 2 shows the correct recognition rate.

**Table 1**. Subject dependent recognition result

| | Actor1 | | Actor2 | | Actor3 | | Actor4 | | Actor5 | | Actor6 | | Actor7 | | Aver. % | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Point. | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 100 | 100 |
| Raise. | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 100 | 95.2 |
| Wave | 6 | 3 | 6 | 3 | 6 | 3 | 5 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 97.6 | 100 |
| Touch | 6 | 2 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 100 | 95.2 |
| Comm | 6 | 2 | 5 | 3 | 6 | 3 | 6 | 3 | 5 | 2 | 6 | 3 | 6 | 3 | 95.2 | 90.5 |
| Bow | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 100 | 100 |
| Pick | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 100 | 100 |
| Kick | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 100 | 100 |
| Walk | 6 | 2 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 6 | 3 | 100 | 95.2 |

**Table 2.** Subject independent recognition result

| | Train set(%) | Test set(%) |
|---|---|---|
| Point. | 97.8 | 100 |
| Raise. | 100 | 100 |
| Wave | 95.6 | 88.9 |
| Touch | 95.6 | 94.4 |
| Comm | 88.9 | 83.3 |
| Bow | 100 | 100 |
| Pick | 100 | 94.4 |
| Kick | 100 | 100 |
| Walk | 100 | 94.4 |

**Table 3**. Comparison with view variant action recognition methods

| | Envelop Shape (%) | | Horizontal projection of silhouettes (%) | | Motion Feature [19] (%) | |
|---|---|---|---|---|---|---|
| Point. | 100 | 94.4 | 44.4 | 88.9 | 38.8 | 100 |
| Raise. | 100 | 100 | 61.1 | 94.4 | 33.3 | 94.4 |
| Wave | 88.9 | 94.4 | 33.3 | 61.1 | 55.5 | 94.4 |
| Touch | 94.4 | 88.9 | 38.8 | 83.3 | 22.2 | 100 |
| Comm | 83.3 | 83.3 | 27.7 | 61.1 | 27.7. | 88.9 |
| Bow | 100 | 100 | 61.1 | 94.4 | 38.8 | 94.4 |
| Pick | 94.4 | 100 | 38.8 | 72.2 | 33.3 | 83.3 |
| Kick | 100 | 94.4 | 55.5 | 83.3 | 38.8 | 88.9 |
| Walk | 94.4 | 94.4 | 33.3 | 72.2 | 44.4 | 100 |

In order to make our approach on view invariant action recognition more convincing, we also tried some comparison experiments. In table 3, we give our results compared with view dependent methods. The table's first row shows three kinds of methods we used. The first method is our proposed approach, "Envelop Shape". As the second method, we use vector of horizontal projection of silhouettes (that is, the "x" in expression (2)) as input, which is view variant. As the third method, we refer to [19] which uses motion features and is a view dependent method. Below

each method, there are two sub columns. The first column gives recognition rates of subject independent action recognition for any view points. The second column gives average recognition rates at a specified view point. We can see that in view independent scenario, only “Envelop Shape” method performs well, however the other two methods perform poor since they are view dependent.

## 4 Conclusion

With the help of “Envelop Shape” representation, we set up an action recognition system despite of viewpoint variance. Experiment shows that with the help of Envelop Shape representation, our system has achieved a high correct recognition rate in the case of free view point actions. Results indicate that "Envelop Shape" representation contains enough discriminating features even for subject-independent action recognition.

“Envelop Shape” representation is view insensitive, and compared to previous approaches, it is easier to acquire and has more abundant information. It does not need any meaningful feature detection or point correspondence, which as we know is often difficult to get and sensitive to errors. However as a view insensitive representation, it loses some view variant information sometimes important for certain action recognition. For example, we can not distinguish left or right hand’s movement only with this representation. Some view variant information may help for solving this kind of problem. How to combine this representation and other view variant information? It is further work to accomplish.

**Acknowledgements**

This work is supported by National Science Foundation of China under grant No 60673189 and No 60433030.

## References

1. C. Cedras, M. Shah, Motion-based recognition: a survey, Image and Vision Computing, 13 (2) (1995) 129-155.
2. J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding, 73 (3) (1999) 428-440
3. T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding, 81 (3) (2001) 231-268.
4. Liang Wang, Weiming Hu, Tieniu Tan, Recent Developments in Human Motion Analysis, Pattern Recognition, Vol. 36, No. 3, pp.585-601, 2003
5. M. Leo,T. D’Orazio,P. Spagnolo,International Multimedia Conference,Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks,(2004),124-130
6. Liang Wang, Tieniu Tan, Huazhong Ning, Weiming Hu, Silhouette analysis-based gait recognition for human identification, IEEE Trans on Pattern Analysis and Machine Intelligence,Vol25,No 12,pp 1505 - 1518,Dec. 2003
7. R. O. Duda, P.E. Hart, D. G. Stock, Pattern Classification, pp 11

8. L.W. Campbell, D.A. Becker, A. Azarbayejani, A.F. Bobick,and A. Pentland, Invariant Features for 3D Gesture Recognition, in Proceedings of International Conference on Automatic Face and Gesture Recognition, pp. 157-162, 1996.
9. Steven M. Seitz1 and Charles R. Dyer1,View-Invariant Analysis of Cyclic Motion, International Journal of Computer Vision, 1997
10. Cen Rao, A. Yilmaz,M. Shah,View-Invariant Representation And Recognition of Actions, International Journal of Computer Vision, Vol. 50, Issue 2, 2002
11. Cen Rao,M. Shah,T. S. Mahmood,Action Rectionition based onView Invariant Spatio-temporal Analysis ,ACM Multimedia 2003, Nov 2-8, Berkeley, CA USA
12. Parameswaran, V., Chellappa, R, Using 2D Projective Invariance for Human Action Recognition, International Journal of Computer Vision,2005
13. Parameswaran, V., Chellappa, R., Human Action Recognition Using Mutual Invariants, Computer Vision and Image Understanding, 2005
14. D. Weinland, R. Ronfard, E. Boyer, Free Viewpoint Action Recognition using Motion History Volumes, Computer Vision and Image Understanding, 2006
15. Wren C, Azarbayejani A, Darrell T and Pentland A. Pfinder: real-time tracking of the human body. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19 (7): 780-785.
16. J.Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequetial images using hidden markov model. In Proc. 1992 IEEE Conf. on Computer Vision and Pattern Rec., pages 379¨C385. IEEE Press, 1992.
17. Brand M, Oliver N, Pentland A. Coupled hidden Markov models for complex action recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Puerto Rico,1997
18. Y.A. Ivanov, A.F. Bobick, Recognition of visual activities and interactions by stochastic parsing, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 22,No 8, Aug. 2000 pp:852 – 872
19. O. Masoud and N. Papanikolopoulos, A method for human action recognition, Image and Vision Computing, Vol. 21, 2003, pp.729-743.