VIEW-INVARIANT ACTION RECOGNITION USING CROSS RATIOS ACROSS FRAMES

Yeyin Zhang, Kaiqi Huang, Yongzhen Huang and Tieniu Tan

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{ yyzhang, kqhuang, yzhuang, tnt}@nlpr.ia.ac.cn

ABSTRACT

We present a new method of computing invariants in videos captured from different views to achieve view-invariant action recognition. To avoid the constraints of collinearity or coplanarity of image points for constructing invariants, we consider several neighboring frames to compute cross ratios, namely cross ratios across frames ($CRAF$), as our invariant representation of action. For every five points sampled with different intervals from the trajectories of action, we construct a pair of cross ratios (CRs). Afterwards, we transform the CRs to histograms as the feature vectors for classification. Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods in effectiveness and stability.

Index Terms— view-invariance, action recognition, cross ratio

1. INTRODUCTION

Human action recognition has gained much attention in the past few years in the applications like visual surveillance, human-computer interfaces, video annotation, content based video retrieval, etc. However, applications remain limited due to difficulties in real circumstances.

View-invariance is a challenging problem in human action recognition. Several kinds of methods concerning that problem have been proposed recently. Firstly, 3D reconstruction techniques could provide the most reliable view independent representation of actions. In [1] and [2], images recorded by multiple calibrated cameras are projected back to 3D visual hulls of the body in different poses. But the high cost of this method limits its practical applications. Secondly, the epipolar geometric relations in multiple view geometry lead to some constraints between image points in different views. For example, [3] uses fundamental ratios, which are the ratios of fundamental matrix and proved to be invariant to viewpoint, to represent pose transitions in a model based method. The problem is that manually labeling of joints of the human body is required to find those triplets of points used for homography calculation. Thirdly, learning methods map motion representation to viewpoint, like [4, 5]. These methods do not need to extract view-invariant features from images. Nevertheless, they implicitly assume the mapping should satisfy the underlying models with empirical priors and are not clear of which aspect of their representation of action accounts for the output. Lastly, people also try to construct invariants from images. [6] applies a spatio-temporal curvature of 2D trajectory of hand to capture dramatic changes of motion. The curvature is view-invariant but with the second derivative, which degrades signal-to-noise ratio unless the curve is smooth enough. The method in [7] assumes that there is a moment in an action when some of the joints of the body are coplanar, namely canonical pose. The application is also limited since it's hard to detect such a canonical pose in videos automatically.

Invariants are usually used for object recognition to tackle the problem of projective distortion caused by viewpoint variations. In view-invariant action recognition, people are more inclined to model based recognition methods. These methods evaluate the fitness between image points and the predefined 3D models. But it's difficult to detect such image points that satisfy the specific geometric configuration required to get the desired invariants. For example, to get a cross ratio as an invariant, image points are required to be collinear or coplanar in the original 3D space before projection.

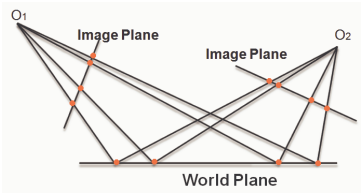

Fig. 1. The sets of four points with identical cross ratio under projective transformations

Cross ratio is the most common invariants. As is shown in Fig.1, the sets of four collinear points with the same permutation lying on different planes form cross ratios with the same value.

To avoid the constraints of collinearity or coplanarity of image points when constructing invariants from an image, we calculate invariants across neighboring frames, rather than from a single image.

We generalize the cross ratio of four collinear points to cross ratios across frames ($CRAF$) using five neighboring coplanar points sampled from trajectories of actions. Under the assumption of coplanarity of neighboring points on the trajectory, the proposed method does not need human body model and manually labeling of image points. Experimental results show that our method outperforms others' with high effectiveness and stability.

The paper is organized as follows: In Section 2, we elaborate the algorithm of cross ratios across frames ($CRAF$). Section 3 details empirical studies on public database and analyzes the stability of $CRAF$. Finally, conclusions are made in Section 4.

2. INVARIANTS ACROSS FRAMES

In our work, we assume trajectories of several key joints on the body, like hand, foot or head, could be obtained by feature tracking techniques. Once we get trajectories from image sequences, we could construct a pair of cross ratios for every five points sampled from the trajectories. Afterwards, pairs of cross ratios are transformed to histograms as the feature vectors in SVM classification. Detailed description of our method is provided in the following subsections.

2.1. Cross ratios across frames

Geometric invariants capture invariant information of a geometric configuration under a class of transformations. Group theory gives us theoretical foundation for constructing invariants [8]. In computer vision applications, we use invariants as our view-invariant representation because such invariants could be measured directly from images without knowing the orientation and position of the camera.

For the difficulty of detecting groups of collinear points in a single image, we construct invariants across frames, that's to say, we use several neighboring frames in a video to compute invariants as our view-invariant representation. The only assumption we should make is the coplanarity of neighboring points on the trajectory.

Cross ratio is invariant to projective transformations. It is defined as:

$$[X_1, X_2, X_3, X_4] = \frac{(X_1 - X_3)(X_2 - X_4)}{(X_1 - X_4)(X_2 - X_3)} \tag{1}$$

Here X_1, X_2, X_3 and X_4 represent a set of four collinear points and the value of $[X_1, X_2, X_3, X_4]$ is preserved by projective transformations.

The precondition of collinearity makes the application of cross ratio of four collinear points limited. So we make a generalization by constructing a pair of cross ratios in the same way as in [8]. As illustrated in Fig.2, suppose we have got a trajectory T and there are five points $(X_1, X_2, X_3, X_4, X_5)$ which are approximately coplanar, we use these 5 points on the trajectory to generate two groups of four collinear points,

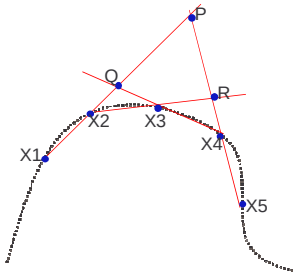

Fig. 2. 5 points to construct a pair of cross ratio

(X_1, X_2, P, Q) and (X_5, X_4, P, R). With the two groups of collinear points, we compute their cross ratios respectively and denote them as CR_1 and CR_2. Thus, we get the view-invariant representation of these five points as follows:

$$T(X_1, X_2, X_3, X_4, X_5) \mapsto I(CR_1, CR_2) \tag{2}$$

where the $I(CR_1, CR_2)$ denotes our view invariant representation of five trajectory points. The only precondition of this generalization is the coplanarity of the five points. Empirical tests in Section 3 show that precondition is satisfied in most real cases.

Computing CR_1 and CR_2 is straightforward as long as the coordinates of the five points on image plane are known. Here we use formulae that have been proven in higher geometry:

$$CR_1(X_1, X_2, P, Q) = \frac{(|X_1X_4| + |X_4X_5| + |X_5X_1|)}{(|X_2X_4| + |X_4X_5| + |X_5X_2|)} \times \frac{(|X_2X_4| + |X_4X_3| + |X_3X_2|)}{(|X_1X_4| + |X_4X_3| + |X_3X_1|)} \tag{3}$$

$$CR_2(X_5, X_4, P, R) = \frac{(|X_5X_2| + |X_2X_1| + |X_1X_5|)}{(|X_5X_2| + |X_2X_3| + |X_3X_5|)} \times \frac{(|X_4X_2| + |X_2X_3| + |X_3X_4|)}{(|X_4X_2| + |X_2X_1| + |X_1X_4|)} \tag{4}$$

where $|X_iX_j|$ is the determinant of the 2×2 matrix $[X_iX_j]$.

Degenerated groups of points might appear while computing CRs. For example, the line defined by X_1 and X_2 is parallel to the line defined by X_4 and X_5, or X_2, X_3 and X_4 are collinear. In these cases, we either assign a fixed number to CR relatively large or just ignore them. Since most of the sampled points are in general position, the degenerated groups do not affect the outputs of the algorithm.

2.2. $CRAF$ histograms

For each trajectory, we get a sequence of pairs of cross ratios. These CRs are voted into bins to form a histogram as the representation of the feature vector for classification. In detail, the value of each histogram bin is defined as:

$$H(i) = \frac{1}{C} \sum_{i=1}^{C} X_i,$$

$$and \quad X_i = \begin{cases} 1 & \text{if } CR_i \in [b(i), b(i+1)) \\ 0 & \text{else} \end{cases} \tag{5}$$

where C is the count of CRs, and $b(i)$ and $b(i+1)$ correspond to the lower-bound and upper-bound of the i^{th} bin of the histogram.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In our experiment, we use CMU Motion Capture (Mocap) Database[1] to get trajectories. MoCap database records 3D position information captured from sensors on the body. After projection onto image planes, we could get 2D trajectories in different views.

To make a comparison with the state of the art, our experiment is conducted under the same condition with [3].

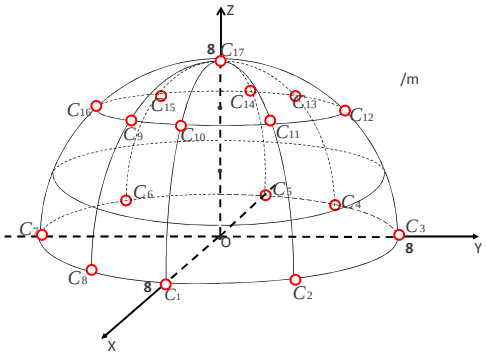

Fig. 3. Camera position distribution

In the projection process, there are seventeen synthesized cameras uniformly distributed around a hemisphere. The distribution of the cameras is depicted in Fig.3. All the actions are performed around the center within the hemisphere. We project 3D data onto images of each viewpoint with the focal length randomly chosen in a range of 1000±300mm. Here is an example of projected trajectories of hand shown in Fig.4, which illustrates the action of jump in each viewpoint with varying appearance caused by projective distortions.

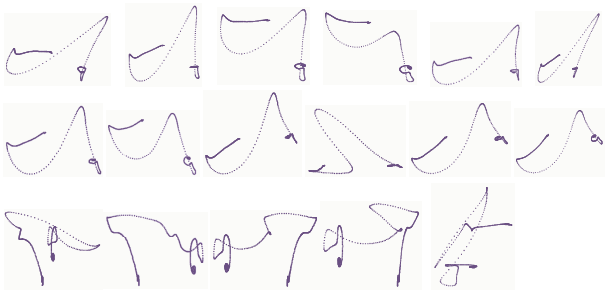

Fig. 4. The projected trajectories of hand of each viewpoint

We select 5 classes of actions, climb, jump, run, swing and walk, from the database to make our tests. For each action, we get trajectories of head, left hand and left foot of the subject. For every neighboring five points on the trajectory, we compute a pair of CRs by Eq.3 and Eq.4. We transform the CRs of each action to histograms as the view-invariant features of the action.

3.1. Recognition results

After projection, We get 200 trajectories of each viewpoint, specifically 12 sequences for climb, 57 sequences for jump, 41 sequences for run, 10 sequences for swing and 80 sequences for walk. The data provided is unbalanced, so weighted training strategy is applied in the training process. We use support vector machine (SVM) as the classifier. In SVM training, the kernel parameters are optimized by way of grid search. We train one model for each viewpoint and test it on the other viewpoints. The output of each viewpoint is the one with the highest score.

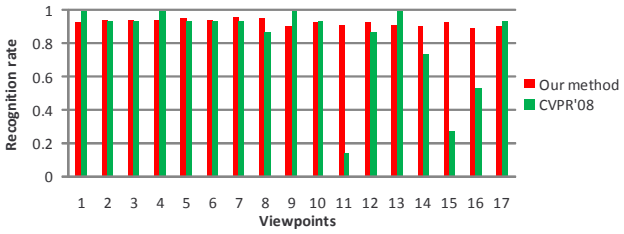

Fig. 5. Recognition rate in different views

The performance is shown in Fig.5. Though the recognition rate of some views is a little lower than that in [3], the average accuracy is about 92.38%, which is much higher compared to 81.60% in [3], demonstrating high stability over the seventeen viewpoints. Besides, unlike only 15 samples were tested in each viewpoint in [3], we used 200 samples for testing in each view, so our output is more persuasive than [3].

In addition, we give the accuracy of each model in each viewpoint in Table 1. Each model in the first column is trained under its corresponding viewpoint and tested under the other viewpoints.

The recognition rate of each model is quite stable in each viewpoint, indicating high robustness to varying viewpoints.

3.2. Stability of $CRAF$ over different sampling intervals

Theoretically, cross ratios of five coplanar points in general position remain the same under projective transformations.

Since we have assumed that the five points used to compute CRs are approximately coplanar, we evaluate the variance of the CRs of groups of neighboring five points at different sampling rate. The variance and mean curve with respect to different sampling rate is shown in Fig.6.

The mean value of the CRs is around 0.6. As we can see in the figure, the variance is negligible compared to the mean value when the sampling rate is above 25Hz, which is

[1]CMU MoCap database: mocap.cs.cmu.edu

Table 1. Recognition rate in each view of each model

	View1	View2	View3	View4	View5	View6	View7	View8	View9	View10	View11	View12	View13	View14	View15	View16	View17	Average
Model_1	1	0.852	0.82	0.809	0.948	0.841	0.811	0.826	0.887	0.813	0.794	0.824	0.878	0.844	0.874	0.837	0.865	0.854294118
Model_2	0.854	1	0.913	0.872	0.846	0.937	0.9	0.9	0.857	0.898	0.883	0.9	0.859	0.813	0.865	0.781	0.844	0.877764706
Model_3	0.805	0.909	1	0.909	0.787	0.93	0.956	0.917	0.822	0.896	0.88	0.893	0.859	0.768	0.852	0.751	0.841	0.869117647
Model_4	0.829	0.915	0.932	1	0.831	0.909	0.926	0.945	0.846	0.911	0.893	0.902	0.852	0.79	0.861	0.781	0.835	0.879882353
Model_5	0.922	0.852	0.807	0.824	1	0.824	0.816	0.807	0.902	0.829	0.796	0.816	0.874	0.887	0.876	0.872	0.854	0.856352941
Model_6	0.824	0.937	0.924	0.889	0.818	1	0.919	0.911	0.865	0.904	0.891	0.902	0.844	0.781	0.85	0.787	0.848	0.876117647
Model_7	0.813	0.904	0.939	0.893	0.8	0.937	1	0.924	0.839	0.896	0.893	0.919	0.861	0.768	0.82	0.766	0.854	0.872117647
Model_8	0.818	0.911	0.913	0.941	0.813	0.906	0.932	1	0.85	0.891	0.896	0.904	0.865	0.772	0.848	0.766	0.837	0.874294118
Model_9	0.872	0.816	0.811	0.826	0.865	0.82	0.818	0.816	1	0.846	0.82	0.844	0.891	0.796	0.824	0.82	0.904	0.846411765
Model_10	0.813	0.88	0.876	0.852	0.796	0.88	0.859	0.861	0.867	1	0.909	0.906	0.861	0.777	0.816	0.748	0.867	0.856941176
Model_11	0.803	0.867	0.885	0.872	0.79	0.911	0.893	0.891	0.857	0.926	1	0.924	0.859	0.753	0.803	0.742	0.857	0.860764706
Model_12	0.79	0.854	0.872	0.861	0.794	0.876	0.861	0.859	0.865	0.898	0.913	1	0.88	0.757	0.796	0.746	0.852	0.851411765
Model_13	0.865	0.826	0.82	0.857	0.846	0.841	0.844	0.829	0.874	0.863	0.857	0.861	1	0.798	0.835	0.783	0.891	0.852352941
Model_14	0.896	0.835	0.757	0.816	0.911	0.79	0.744	0.77	0.848	0.755	0.727	0.753	0.852	1	0.915	0.887	0.807	0.827235294
Model_15	0.896	0.898	0.857	0.885	0.891	0.85	0.846	0.844	0.872	0.841	0.826	0.857	0.88	0.896	1	0.883	0.848	0.874705882
Model_16	0.896	0.831	0.807	0.816	0.902	0.813	0.777	0.781	0.867	0.77	0.761	0.787	0.859	0.904	0.926	1	0.844	0.843588235
Model_17	0.839	0.796	0.79	0.813	0.841	0.813	0.8	0.805	0.9	0.835	0.837	0.857	0.911	0.79	0.816	0.794	1	0.837470588

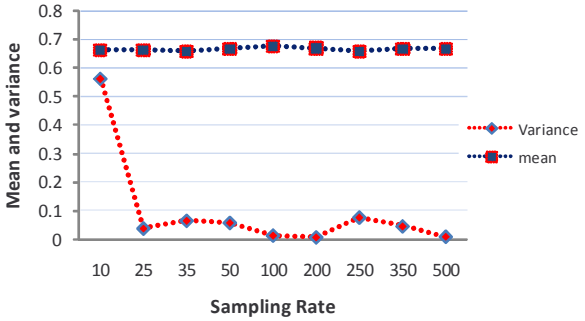

Fig. 6. Mean and variance of CRs in different viewpoints

to say, the calculated CR is stable as long as the frame rate is above 25Hz, indicating our approximate coplanar assumption is acceptable under real circumstances.

4. CONCLUSIONS

In this paper, we proposed a method of computing invariants across frames. We made generalizations to cross ratio of four collinear points so that it could be applied to view-invariant representation of actions. In our stability evaluation, points across frames with different sampling intervals produce invariants with tolerable variance in different viewpoints. In classification, the invariants show high robustness to varying viewpoints and sampling intervals.

Acknowledgement

This work is funded by research grants from the National Basic Research Program of China (2004CB318110), the National Science Foundation (60605014, 60875021), the National Natural Science Foundation of China (60736018, 60723005) and National Laboratory of Pattern Recognition (2008NLPRZY-2). The authors also thank the anonymous reviewers for their valuable comments.

5. REFERENCES

[1] Daniel Weinland, Remi Ronfard, and Edmond Boyer, "Free viewpoint action recognition using motion history volumes," *Int. J. Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249 – 257, 2006.

[2] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," *IEEE 11th International Conference on Computer Vision, 2007*.

[3] Yuping Shen and H. Foroosh, "View-invariant action recognition using fundamental ratios," *IEEE Conference on Computer Vision and Pattern Recognition, 2008.*

[4] P. Natarajan and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," *IEEE Conference on Computer Vision and Pattern Recognition, 2008.*

[5] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," *IEEE Conference on Computer Vision and Pattern Recognition, 2008.*

[6] Cen Rao, Alper Yilmaz, and Mubarak Shah, "View-invariant representation and recognition of actions," *Int. J. Comput. Vision*, vol. 50, no. 2, pp. 203–226, 2002.

[7] Vasu Parameswaran and Rama Chellappa, "View invariants for human action recognition," *Proc. IEEE Conf. Computer Vision Pattern Recognition 2003.*, vol. 2, pp. 613.

[8] Joseph L. Mundy and Andrew Zisserman, Eds., *Geometric invariance in computer vision*, MIT Press, Cambridge, MA, USA, 1992.