Action Recognition Based on A Bag of 3D Points

Wanqing Li
Advanced Multimedia Research Lab, ICT Research Institute
University of Wollongong, NSW 2522, Australia
wanqing@uow.edu.au

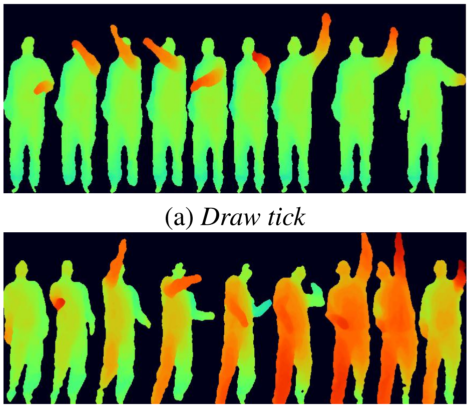
Zhengyou Zhang, Zicheng Liu
Microsoft Research, Redmond
One Microsofr Way, Redmond, WA 98025, US
{zhang,zliu}@microsoft.com

Abstract

This paper presents a method to recognize human actions from sequences of depth maps. Specifically, we employ an **action graph** *to model explicitly the dynamics of the actions and a bag of 3D points to characterize a set of salient postures that correspond to the nodes in the action graph. In addition, we propose a simple, but effective projection based sampling scheme to sample the bag of 3D points from the depth maps. Experimental results have shown that over 90% recognition accuracy were achieved by sampling only about 1% 3D points from the depth maps. Compared to the 2D silhouette based recognition, the recognition errors were halved. In addition, we demonstrate the potential of the bag of points posture model to deal with occlusions through simulation.*

1. Introduction

Recognition of human actions has been an active research topic in computer vision. In the past decade, research has mainly focused on leaning and recognizing actions from video sequences captured from a single camera and rich literature can be found in a wide range of fields including computer vision, pattern recognition, machine leaning and signal processing [14]. Among the different types of visual input, silhouettes and spatio-temporal interest points (STIPs) are most widely used [8]. However, as the imaging technology, especially real-time 3D shape measurement using fringe projection techniques [20], advances, it becomes feasible and also economically sound to capture in real-time not only the conventional two-dimensional (2D) color image sequences, but also the sequences of depth information. This paper studies the recognition of human actions from sequences of depth maps.



(a) *Draw tick*

(b) *Tennis serve*

Figure 1. Examples of the sequences of depth maps for actions: (a) *Draw tick* and (b) *Tennis serve*
.

Figure 1 shows nine depth maps of typical depth sequences for actions *Draw tick* and *Tennis serve*. As seen, the depth map provides additional shape information of the body and is expected to be helpful in distinguishing the actions that may generate similar 2D silhouettes when captured from a single view. However, the depth map also substantially increases the amount of the data to be processed. Effective and efficient use of the depth information is a key to a computationally efficient algorithm for action recognition based on the sequences of depth maps.

As it is well known, the human body is an articulated system of rigid segments connected by joints and human motion is often considered as a continuous evolution of the spatial configuration of the segments or body posture [18]. Given a sequence of depth maps, if the body joints can be reliably extracted and tracked, action recognition can be achieved by using the tracked joint positions. However, tracking of body joints from depth maps is still an unsolved problem. In addition, most real-time depth cameras can only produce coarse and noisy depth maps. This paper aims to develop a technique that does not require joint tracking. We adopt the expandable graphical model [8] to explicitly model the temporal dynamics of the actions and propose to use a bag of 3D points (BOPs) extracted from the depth map to model the postures. An *action graph* is constructed from the training samples to encode all actions to be recognized. Every action is encoded in one or multiple paths in the action graph and each node in the action graph represents a salient posture that is shared by all actions. Our idea is to use a small set of $3D$ points to characterize the 3D shape of each salient postures and to use Gaussian Mixture Model (GMM) to effectively and robustly capture the statistical distribution of the points. To this end, we propose a simple but effective projection based method to sample a very small set of representative 3D points from the depth maps. Experimental results have shown that over 90% recognition accuracies are achieved by only sampling 1% 3D points from the depth maps. Compared to the 2D silhouette based recognition, the recognition errors were halfed. In addition, we have demonstrated the potential of the bag of points posture model to deal with occlusions through simulation.

The rest of the paper is organized as follows. Section 2 gives a short review of the recent advance on silhouette based human action recognition. Section 3 describes the *action graph* based on the concept of bag-of-points. In Section 4, the projection-based 3D points sampling scheme is presented. Experimental results are given in Section 5 and discussion and future extension of the proposed method are presented in Section 6

2. Related work

A rich palette of diverse ideas has been proposed during the past few years on the problem of recognition of human actions by employing different types of visual information. Among them, silhouettes and spatio-temporal interesting points (STIPs) have been most widely used visual features. However, to our best knowledge, there are few reports so far on action recognition using depth maps. Therefore, This section presents a review on the work of silhouette-based action recognition since it is most relevant to the work presented in this paper. Review on other types of visual features, such as STIPs and skeletons, can be found in [12, 11, 14, 8].

Use of silhouettes as input to recognize human actions in general follows the concept that human motion can be considered as a continuous evolution of the body pose and is mainly driven by the background modeling techniques and application of video surveillance. Silhouettes can effectively describe the shape of body poses. Methods proposed in the past for silhouette-based action recognition can be classified into two categories. One is to extract *action descriptors* from the sequences of silhouettes. Conventional classifiers are used for the recognition. In this approach, the action descriptors are supposed to capture both spatial and temporal characteristics of the actions. For instance, in the early work by Bobick and Davis [2], the silhouettes are stacked together to form a motion energy images (MEI) and motion history images (MHI). Seven Hu moments [6] are extracted from both MEI and MHI to serve as action descriptors. Action recognition is based on the Mahalanobis distance between each moment descriptor of the known actions and the input one. Meng [10] extended the MEI and MHI into a hierarchical form and used a Support Vector Machine (SVM) to recognize the actions. In the method proposed by Chen et al. [3], star figure models were fitted to silhouettes to capture the five extremities of the shape that correspond to the arms, legs and head. Gaussian mixture models were used to model the spatial distribution of the five extremities over the period of an action, ignoring the temporal order of the silhouettes in the action sequence. Davis and Yyagi [5] also used GMM to capture the distribution of the moments of a silhouette sequence.

The other category of the methods is to extract features from each silhouette and model the dynamics of the action explicitly, often using statistical models such as hidden Markov models(HMM), graphical models (GM) and conditional random fields (CRF). The feature extracted from each silhouette is to capture the shape of the body and possible local motion. For instance, Divis and Tyagi [5] used moments to describe shapes of a silhouette and continuous HMM to model the dynamics. In [7], Kellokumpu et al. chose Fourier shape descriptors and classified the postures into a finite number of clusters. Discrete HMM are then used to model the dynamics of the actions where the posture clusters are considered to be the discrete symbols emitted from the hidden states. Sminchisescu et al. [13] relaxed the HMM assumption of conditional independence of observations given the actions by adopting the CRF model. Zhang and Gong [19] extracted the spectrum of the chain-code of the silhouettes and motion moment as the feature and proposed a modified hidden conditional random field (HCRF) to model the dynamics of the actions. Veerarahavan, et al. [15] adopted Kendall's representation of shape as shape features and proposed to use autoregressive (AR) model and autoregressive and moving average (ARMA) model to capture the kinematics of the actions. In the work by Wang and Suter [16], Locality Preserving Projection (LPP) was employed to learn a subspace to describe the postures and DTW and temporal Huasdorff distance to classify the actions in the subspace. Colombo, et al. [4] proposed to find the subspace for each type of actions through principal component analysis (PCA). Li *et al.* [8] extracted the holistic motion and 32 2D points from the contour of the silhouettes as the feature and constructed an expandable graphical model to represent the postures and dynamics of the postures explicitly. In the most recent work, Yu and Aggarwal [17] proposed to use the five extremities extracted from the silhouettes as a semantic posture presentation. The distribution of the five extremities represented as a histogram similar to Belongies's shape-context [1] was used as the feature for the construction of a HMM for each action.

From the extensive work on silhouette based action recognition, we have observed that silhouettes contain rich shape information of the body. However, most shape information is carried by the points on the contour rather than the points inside the silhouettes. Even the points on the contour appear to have redundancy as the body shape can be well described by five extremities if they are extracted correctly. This observation forms the core concept of the method proposed in this paper and is extended to the 3D case.

3. Action graph based on bag-of-points

Action graph [8] is an effective method to explicitly model the dynamics of human motion based on a set of salient postures shared among the actions. An *action graph* that encodes L actions with M salient postures $\Omega = \{\omega_1, \omega_2, \cdots, \omega_M\}$ can be represented as

$$G = \{\Omega, A, A_1, A_2, \cdots, A_L\}, \tag{1}$$

where each posture serves as a node, $A_k = \{p(\omega_j|\omega_i, \psi_k)\}_{i,j=1:M}^{k=1:L}$ is the transitional probability matrix of the $k'th$ action ψ_k and $A = \{p(\omega_j|\omega_i)\}_{i,j=1}^{M}$ is the global transitional probability matrix of all actions.

An action recognition system that is built upon the action graph can be then described by a quadruplet,

$$\Gamma = (\Omega, \Lambda, G, \Psi), \tag{2}$$

where

$$\begin{aligned}
\Omega &= \{\omega_1, \omega_2, \cdots, \omega_M\} \qquad (3)\\
\Lambda &= \{p(x|\omega_1), p(x|\omega_2), \cdots, p(x|\omega_M)\}\\
G &= (\Omega, A, A_1, A_2, \cdots, A_L)\\
\Psi &= (\psi_1, \psi_2, \cdots, \psi_L).
\end{aligned}$$

$p(x|\omega_i)$ is the probability of an instantaneous observation x realized from the salient posture ω_i; ψ_i denotes the $i'th$ trained action. Given such a system Γ and a sequence of observations $X = \{x_1, x_2, \cdots, x_n\}$, five different decoding schemes [8] can be used to recognize the most likely action that produces X. In this paper, we adopt bi-gram with maximum likelihood decoding (BMLD) scheme since it requires least computing resources. More details can be found in [8].

One of the key components in training the system Γ is to construct an action graph from the training samples. Li *et al.* [9, 8] proposed to cluster all training poses (frames) into M clusters, each cluster representing a salient posture. This method has been successfully applied to silhouette based action recognition where the external contour of the silhouettes and holistic motion were used as the feature to characterize the posture.

In this paper, we propose to describe the salient postures using a bag-of-points (BOPs), $x = \{q_i, i = 1, 2, \cdots, m\}$, where point q_i can be a pixel belonging to a silhouette or a STIP. Notice that each posture may have different number of points and there is no correspondence between the points of different postures. Furthermore, we assume that the distribution of the points for a posture can be approximated by a mixture of Q Gaussian components and these points are statistically independent. Consequently, the posture model can be simply represented as the joint distribution of the points. That is,

$$p(x|\omega) = \prod_{i=1}^{m} \sum_{t=1}^{Q} \pi_t^{\omega} g(q_i, \mu_t^{\omega}, \Sigma_t^{\omega}) \tag{4}$$

where $g(\cdot)$ is a Gaussian function.

In the case that the input is a sequence of depth maps, a simple approach is to consider that x consists of all the 3D points from the depth map since this depth map represents the 3D surface of the body with respect to the viewpoint and adequately captures the shape of the body (or pose). However, such an approach will require huge amount of computation due to the number of 3D points involved and also may be severely interfered by the noise in the depth map. On the other hand, the study of using the extreme points of 2D silhouettes for action recognition has suggested that a small set of representative 3D points sampled from the depth map should be sufficient to capture the shape of the body. In the next section, we present a simple yet effective sampling scheme based on the projections of the depth map.

4. Sampling of 3D Points

Previous research on 2D silhouette-based action recognition has shown that the contour and the extreme points of the contour carries much shape information of the body. A simple way to sample the representative 3D points is to extract the points on the contours of the planar projections of the 3D depth map. Depending on the number of projection planes used, however, the number of points can still be large. Therefore, in this paper we adopt to project the depth map onto the three orthogonal Cartesian planes and further sample a specified number of points at equal distance along the contours of the projections . Figure 2 shows the sampling process which consists of projection, contour sampling and retrieval of the 3D points that are nearest to the sampled 2D points. There may be multiple 3D points that are nearest to a 2D point, our current implementation is to randomly select one of them. Notice that the projections to the $xz-$ and $zy-$ plane can be very coarse due to the resolution of the depth map and interpolation may be required in order to construct these projections. In addition, each projection may have multiple unconnected regions, contours of all regions are sampled.

5. Experimental results

5.1. Dataset and setup

Since no public benchmarking datasets which supplies the sequences of depth maps are available, we collected a dataset that contains twenty actions: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*,*bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pickup & throw*. Each action was performed by seven subjects for three times. The subjects were facing the camera during the performance. The depth maps were captured at about 15 frames per second by a depth camera that acquires the depth through structure infra-red light. The size of the depth map is 640 × 480. Altogether, the dataset has 23797 frames of depth maps for the 4020 action samples.

Notice that the 20 actions were chosen in the context of using the actions to interact with game consoles. They reasonably cover the various movement of arms, legs, torso and their combinations. In addition, if an action is performed by a single arm or leg, the subjects were advised to use their right arm or leg.

Due to the large amount of computation involved in clustering the training samples into salient postures, we divided the 20 actions into three subsets, each having 8 actions. Table 1 lists the three action subsets used in the experiments. The AS1 and AS2 were intended to group actions with similar movement, while AS3 was intended to group complex actions together. For instance, action *hammer* is likely to be confused with *forward punch* and action *pickup & throw* is a composition of *bend* and *high throw* in AS1.

In all the experiments, we first down-sampled the depth map by factor of 2 (for the sake of computation) and then sampled 80 3D points from the down-sampled depth map. The 80 3D points represent about 1% of the depth points in the down-sampled map and about 0.25% of the original depth map. Among the 80 3D points, 45% (i.e.36 points) were alloacted to the xy-projection, 30% (i.e. 24 points) to the zy-projection and 25% (i.e. 20 points) to the zx-projection. In all the experiments reported below, training samples were clustered using the Non-Euclidean Relational Fuzzy (NERF) C-Means and the disimilarity between two depth maps was caluclated as the hausdorf distance between the two sets of the sampled 3D points. Both the number of postures and the number of Gaussian components to model the posture were set to 45.

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

Table 1. The three subsets of actions used in the experiments

	Test One	Test Two	Cross Subject Test
AS1	89.5%	93.4%	72.9%
AS2	89.0%	92.9%	71.9%
AS3	96.3%	96.3%	79.2%
Overall	91.6%	94.2%	74.7%

Table 2. Recognition accuracies of different tests. In test one, $1/3$ of the samples were used as training samples and the rest as testing samples; in test two, $2/3$ samples were used as training samples; and in the cross subject test, half of the subjects were used as training and the rest subjects were used as testing.

5.2. Results

Experiments were conducted using different number of training samples in order to evaluate the performance of the proposed method. Table 2 reports respectively the average recognition accuracies when one third of the samples, two third of the samples and samples from half of subjects were used as training. The results have shown that the 80 3D points have captured the body shape very well given the high recognition accuracies.

Notice that the accuracies in the cross subject test are relatively low. This is probably due to the small number of subjects and also the significant variations of the same action performed by different subjects. During the data collection, subjects were free to choose the style of the actions. For instance, some subjects chose to perform the *hand clap* without stretching their arms whereas others did. It is expected that more subjects are required in order to improve the cross subject performance.

5.3. Simulation of occlusion

One of the advantages using BOPs representation is its potential to deal with occlusion. In this experiment, we divided the depth map into four quadrants as shown in Figure 3. Occlusion was simulated by ignoring among the 80

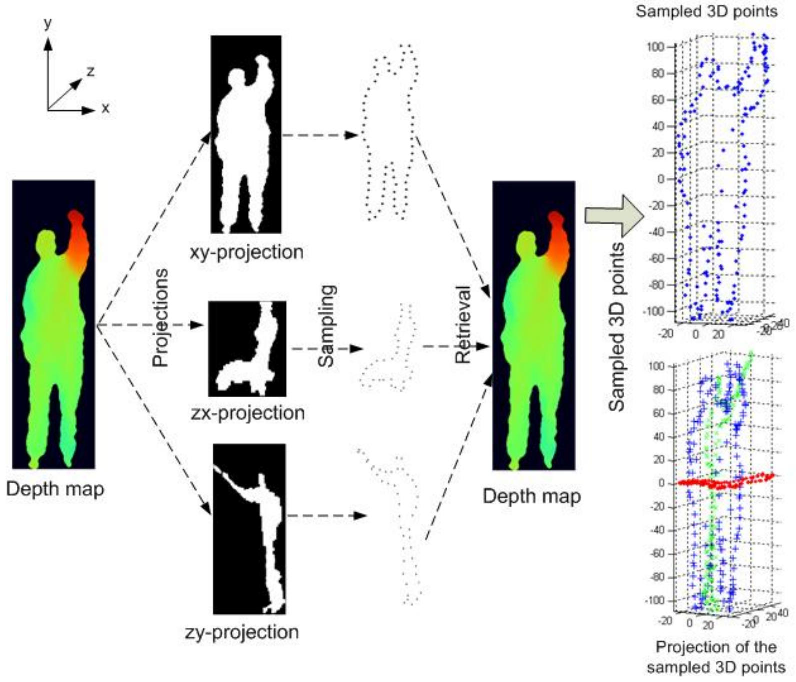


Figure 2. The process of sampling 3D representative points from a depth map.

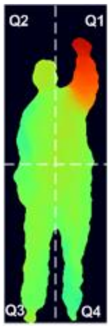


Figure 3. Simulation of occlusion. The depth map is divided into four quadrants.

sampled 3D points those points that fell into the specified occluded quadrants.

We selected the best action graphs trained for the three action subsets respectively and ran through the simulation. Table 3 shows the recognition accuracies after the specified quadrants were omitted during the test. The results have shown that, unless the critical quadrants, for instance, Q1 and Q2 for AS1, were missing, the performance drop was relatively small.

5.4. Comparison to 2D silhouettes

Experiments were also conducted to evaluate the recognition accuracy of using 2D silhouettes. We extracted the 2D silhouettes using xy-projections and 80 points were sampled from the contour of each 2D silhouette. With the same number of postures, the same number of Gaussian components for each posture and the same number of training samples, Table 4 shows the recognition accuracies based on the 2D silhouettes. Compared to the 3D results shown in Table 2, it can be seen that error rates are doubled. This is probably because the action samples are captured from the front viewpoint and depth plays a key role in describing the shape of the posture in many of the 20 actions, such as *hammer*, *forward punch*, *forward kick* and *draw tick*. For example, a subject moves his hand in front of his body (not to the side) to perform the action *draw tick*, the tip of the hand will likely be a sampled point in the 3D representation, but it would not be selected in the 2D representation

Occluded Quadrants	AS1	AS2	AS3
none	98.2%	98.2%	100%
Q1	80.4%	75.0%	88.9%
Q2	91.1%	91.1%	91.8%
Q3	96.4%	94.7%	96.3%
Q4	98.2%	92.9%	94.5%
Q1+Q2	62.5%	95.5%	85.2%
Q3+Q4	91.1%	94.7%	96.3%
Q2+Q3	82.2%	89.3%	95.5%
Q1+Q4	91.1%	94.7%	96.3%

Table 3. Recognition accuracies for the simulated occlusion.

	Test One	Test Two	Cross Subject Test
AS1	79.5%	81.3%	36.3%
AS2	82.2%	88.7%	48.9%
AS3	83.3%	89.5%	45.8%
Overall	81.7%	86.5%	43.7%

Table 4. Recognition accuracies based on 2D silhouettes

because it is not on the silhouette contour.

6. Discussion and Future work

This paper presents a study on recognizing human actions from sequences of depth maps. We have employed the concept of BOPs in the expandable graphical model framework to construct the action graph to encode the actions. Each node of the action graph which represents a salient postures is described by a small set of of representative 3D points sampled from the depth maps. Experimental results have clearly shown the promising performance of the proposed method and also the advantages of using 3D points over the 2D silhouettes. Our plan is to collect more data from more subjects and run the proposed method through all 20 actions at same time instead of dividing them into three subsets in the near future.

Arguably, those actions that were performed by a single arm can be recognized by the recognition methods designed for arm gestures. In these methods, the arm is often tracked and recognition is often based on its trajectory. However, our method provides a unified framework for the actions performed by both single or multiple parts of the body and, most importantly, it does not require tracking.

Viewpoint variance is an important, but challenging problem in action recognition. We note that the current sampling scheme is view dependent. Use of the 3D points offers a potential to extend the proposed approach into a view invariant one. This is currently being studied by the authors.

Acknowledgment

Wanqing Li would like to thank Communication and Collaboration Systems Group for offering the support and opportunity to develop the work at MSR, Redmond.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. PAMI*, 24(4):509–522, 2002. 3

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. PAMI*, 23(3):257–267, 2001. 2

[3] D.-Y. Chen, H.-Y. M. Liao, and S.-W. Shih. Humnn action reocgnition using 2-D spatio-temporal templates. In *Proc ICME*, pages 667–670, 2007. 2

[4] C. Colombo, D. Comanducci, and A. Del Bimbo. Compact representation and probabilistic classification of human actions in videos. In *Proc IEEE Conf. Advanced Video and Signal Based Surveillance (AVSS)*, pages 342–346, 2007. 2

[5] J. W. Davis and A. Tyagi. Minimal-latency human action recognition using reliable-inference. *Image and Vision Computing*, 24(5):455–472, 2006. 2

[6] M. Hu. Visual pattern recognition by moment invariants. *IRE Trans. Information Theory*, 8(2):179–187, 1962. 2

[7] V. Kellokumpu, M. Pietikainen, and J. Heikkila. Human activity recognition using sequences of postures. In *Proc IAPR Conf. Machine Vision Applications*, pages 570–573, 2005. 2

[8] W. Li, Z. Zhang, and Z. Liu. Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1499–1510, 2008. 1, 2, 3

[9] W. Li, Z. Zhang, and Z. Liu. Graphical modeling and decoding of human actions. In *Proc. IEEE MMSP*, 2008. 3

[10] H. Meng, N. Pears, and C. Bailey. A human action recognition system for embedded computer vision application. In *Proc. CVPR*, 2007. 2

[11] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001. 2

[12] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and Image Understanding*, 104(2-3):90–126, 2006. 2

[13] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *Proc ICCV*, volume 2, pages 808–815, 2005. 2

[14] P. Turaga, R. Chellapa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008. 1, 2

[15] A. Veeraraghavan, A. R. Chowdhury, and R. Chellappa. Role of shape and kinematics in human movement analysis. In *IEEE COnf. on CVPR*, volume 1, pages 730–737, 2004. 2

[16] L. Wang and D. Suter. Learning and matching of dynamic shape manifolds for human action recognition. *IEEE Trans. Image Processing*, 16:1646–1661, 2007. 2

[17] E. Yu and J. K. Aggarwal. Human action recognition with extremities as semantic posture representation. In *Proc CVPR*, 2009. 3

[18] V. M. Zatsiorsky. *Kinematics of Human Motion*. Human Kinetics Publishers, 1997. 1

[19] J. Zhang and S. Gong. Action categoration with modified hidden conditional random field. *Pattern Recognition*, 43:197–203, 2010. 2

[20] S. Zhang. Recent progresses on real-time 3D shape measurement using digital fringe projection techniques. *Optics and lasers in engineering*, 48(2):149–158, 2010. 1