Action Categorization based on Arm Pose Modeling

Chongguo Li and Nelson H. C. Yung

Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China

Keywords: Action Categorization, Arm Pose Modeling, Graphical Model, Maximum a Posteriori.

Abstract: This paper proposes a novel method to categorize human action based on arm pose modeling. Traditionally, human action categorization relies much on the extracted features from video or images. In this research, we exploit the relationship between action categorization and arm pose modeling, which can be visualized in a probabilistic graphical model. Given visual observations, they can be estimated by maximum a posteriori (MAP) in that arm poses are first estimated under the hypothesis of action category by dynamic programming, and then action category hypothesis is validated by soft-max model based on the estimated arm poses. The prior distribution of each action is estimated by a semi-parametric estimator in advance, and pixel-based dense features including LBP, SIFT, colour-SIFT, and texton are utilized to enhance the likelihood computation by the Joint Adaboosting algorithm. The proposed method has been evaluated on images of walking, waving and jogging from the HumanEva-I dataset. It is found to have arm pose modeling performance better than the method of mixtures of parts, and action categorization success rate of 96.69%.

1 INTRODUCTION

Human action categorization from visual observations leads to answering the question of “what is the person doing?”. Traditionally, action categorization involves human pose estimation and action recognition, of which they are always treated separately (Moeslund et al., 2011). In fact, action and pose are often perceived simultaneously. A research (Yao et al., 2011) addressed the question of whether pose estimation is useful for action categorization, and their experiments confirmed that action categorization indeed can benefit from pose estimation. On the other hand, if action category is incorporated, human pose estimation can be improved significantly (Li and Yung, 2012), because action information helps deduce possible poses and narrows the pose searching space.

By and large, arm pose as a subset of human pose is far more representative of the action taken than poses by other body parts and therefore dominates the process of action categorization. It is well-known that there are general arm poses for different actions although individual interpretation may be somewhat different. The difference in interpretation may be due to individual style, body posture, as well as action targets. In spite of the differences, arm pose of a specific action is usually constrained by its prior, which defines the movement trend of the action. If the action trend matches the prior, then deviation in other details is tolerable while the action is recognizable. As such, arm pose modeling and action categorization are complementary from a visual perception point of view.

In this research, arm pose modeling and action categorization are investigated as two aspects of the same question. It can be seen as arm pose modeling estimates arm positions while action categorization assigns the sequence of movements to the most likely action category, and action priors in turn refine the estimated arm poses. The relations between action category, arm poses and visual data can be depicted as a hierarchical graphical model in which the action category is treated as the topic variable, the arm pose modeling is the latent variable, and the visual data is the observed variable. The topic variable and the latent variable are the objects to be estimated based on visual observations with the help of pre-learned action priors. Given the visual observations, the topic variable and the latent variable that maximize a posteriori (MAP). In order to infer the MAP efficiently, the topic variable and the latent variable are observed alternatively. The best fitted action category and the corresponding arm poses denote the final results.

The main contributions of the proposed method are as follows. Firstly, a graphical model is proposed for action categorization and arm pose modeling, and two stages inference are adopted. It incorporates the visual evidence of individual arm parts and their prior

Li C. and Yung N..
Action Categorization based on Arm Pose Modeling.
DOI: 10.5220/0004671500390047
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 39-47
ISBN: 978-989-758-004-8

distributions of different actions. Secondly, multiple dense features are used to enhance arm part likelihood, and semi-parametric density estimation is used for arm pose of actions. Thirdly, it has been evaluated on the HumanEva-I dataset and shown significant improvement over the method of mixtures of parts (Yang and Ramanan, 2011), as well as 96.69% success rate on action recognition.

2 RELATED WORKS

Broadly, there are two main directions in this research area (Moeslund et al., 2011): first, discriminative approach treats action categorization as a specific labeling method; and second, generative model based approach uses probabilistic models to capture the inherent relations between the observation variables and hidden states of human action.

Discriminative approach for human action is a classification solution for labeling action and the classifiers are learned from training datasets. (Schuldt et al., 2004) proposed a SVM classification schemes for human action recognition which adopts a local space-time feature (Laptev, 2005) to capture local events. It has an average 86.6% recognition accuracy on the KTH dataset. A codebook based on estimated dynamic pose has been used for action categorization by SVM classifiers (Xu et al., 2012). It has 91.2% and 81.33% average accuracies on the KTH (Schuldt et al., 2004) and UCF sports datasets (Rodriguez et al., 2008) respectively. But it heavily depends on the accuracy of pose estimation (Yang and Ramanan, 2011). Action bank (Sadanand and Corso, 2012), a high-level representation of activity in video with many individual action detectors, is used as features for a linear SVM classifier on KTH and UCF sports with 98.2% and 95% average accuracies respectively. But, this approach needs human to select templates. Bag of Poses (BoP) (Gong et al., 2013), inspired by the idea of Bag of Word, uses weak poses to form the action vocabulary and SVM for action recognition. It was evaluated on the dataset of HumanEva-I and IXMAS with 93.9% and 82.2% action recognition rate respectively. (Fathi and Mori, 2008) constructed three levels of classifiers from low-level optical flow features to the final classifier for action categorization by AdaBoost. Its average performances on KTH and Weizmann (Blank et al., 2005) are 90.5% and 99% respectively.

Generative models for action categorization are also called parametric time-series methods, which involve learning probabilistic models for various human actions. (Yamato et al., 1992) proposed an HMM based method to recognize tennis playing actions. They used vector quantification to convert grid-based silhouette mesh features to an observation sequence. In action categorization, the HMM that best matches the observation sequence is chosen as the correct action sequence. Its recognition rate is higher than 90% for six tennis strokes. An extension of HMM combines duration modeling, multi-channel interactions and hierarchical structure into a single model (Natarajan and Nevatia, 2012) to capture the duration of subevent, the interactions among agents, and the inherent hierarchical organization of activities. The overall accuracy rates are 90.6% on a gesture dataset (Elgammal et al., 2003) and around 100% on the Weizmann (Blank et al., 2005). Topic models or hierarchical Bayesian models, such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA), are popular in language processing but also used in action categorization (Niebles et al., 2008). Spatial-temporal words are extracted from space-time interest points (Dollár et al., 2005) and they are assigned to one of many topic models by the MAP of the hierarchical Bayesian model. Its average performance were 83.33% and 90% on the KTH and Weizmann datasets respectively.

In summary, methods for human actions recognition published in the past are reasonably efficient for some human action datasets. Discriminative methods rely on the classification scheme to deal with image features in order to recognize the corresponding human action, while ignore the semantics of human actions completely. Generative models try to describe dependent relations among the related variables of observed image features and action category. Most generative models for human motion categorization are directly based on feature words without any semantics of the human body. They attempt to map visual observations to an action category directly, but ignore the body configuration altogether. This is obviously different from the way we recognize actions. We believe that body configuration is fundamental in action categorization, and the major constituent that defines body configuration is arm pose.

3 PROBLEM FORMULATION

3.1 Graphical Models

If an image is viewed as a document and a video as a set of documents, then the arm pose of an image can be viewed as the word of a document, and the action category simply as the topic of the document. As shown in Figure 1, for every image in a video, the action category z is viewed as a topic variable for arm pose, and the arm pose w is viewed as a word i.e. latent variable for the corresponding visual observation D.

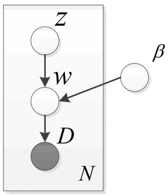


Figure 1: Graphical model for action category z and arm pose w, in which arm pose priors β is fixed and D is the observation.

Furthermore this latent variable also has a Bayesian Network structure. The latent variable probabilities β are the arm pose priors for every action, which we treat as a fixed quantity to be estimated. So this problem is treated as an estimate of the action category z and the arm pose w from the observations D simultaneously. N is the number of image for a video and K is the number of action category that may appear. $z \in \{1,2,\cdots,K\}$ represents the action category, and it indicates an arm pose prior distributions as $\beta_z \in \{\beta_1,\beta_2,\cdots,\beta_K\}$, in which β_K is a set of distributions for arm parts of the action category z. The joint probability of the graphical model as shown in Figure 1 is:

$$\begin{aligned} Pr(D,w,\beta,z) &= Pr(D|w)Pr(w|z,\beta) \\ &= Pr(D|w)Pr(w|\beta_z). \end{aligned} \tag{1}$$

The arm pose priors β can be learned in a supervised manner that the action category z and arm pose w are both observed. During inference, there is an arm pose model based on a hypothesis of action category where z is observed, and an action category validation based on the modeled arm pose where w is derived.

3.2 Bayesian Network of Arm Pose

In Eqt.1, $Pr(D|w)$ and $Pr(w|\beta_z)$ are likelihood and prior of arm pose w respectively. Arm pose w depicts the spatial positions of all the arm parts in a 2D image. As proposed in (Li and Yung, 2012), the arm pose of a person in a 2D image can be defined by seven parameters: shoulder position p, the corresponding orientations φ and scaling factors ρ for upper arm, forearm and hand respectively. From that, the left arm pose θ_L in a 2D image is given by $\theta_L = [p_L, \varphi_{LUA}, \rho_{LUA}, \varphi_{LFA}, \rho_{LFA}, \varphi_{LH}, \rho_{LH}]$, where LUA, LFA and LH stand for left upper arm, left forearm and left hand respectively. The right arm pose θ_R can also be defined in the same way. Therefore, for a person in 2D image, its arm pose parameter is written as $w = [\theta_L; \theta_R]$.

According to the anatomical structure of human body, an arm is attached to the torso via the shoulder and can be viewed as a chain with upper arm, forearm and hand. This anatomical chain also can be mapped to a chain of graphical model. It describes a conditional dependent relation between every arm part. A left arm, for example, its parameters of hand $\theta_{LH} = [\varphi_{LH}, \rho_{LH}]$ depend on the parameters of the forearm $\theta_{LFA} = [\varphi_{LFA}, \rho_{LFA}]$ which depend on the parameters of the upper arm $\theta_{LUA} = [\varphi_{LUA}, \rho_{LUA}]$. The parameter of its shoulder p_L only determines the start position of the upper arm. So there are some conditional independent relations which can be considered as redundancies among the arm component parameters. Figure 2 depicts the dependent relations between all variables.

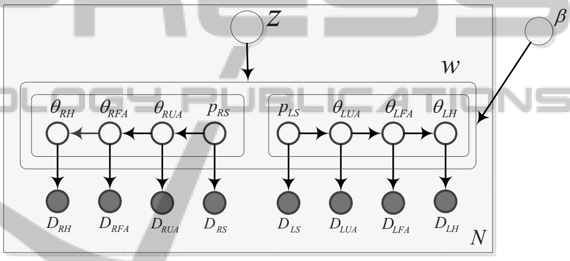


Figure 2: Full graphical model for action category and arm pose modeling.

As such, the joint probability is decomposed as

$$\begin{aligned} Pr(D,w,\beta,z) &= Pr(D|\theta_L,\theta_R)Pr(\theta_L,\theta_R|\beta_z) \\ &= Pr(D|\theta_L)Pr(\theta_L|\beta_z)Pr(D|\theta_R)Pr(\theta_R|\beta_z). \end{aligned} \tag{2}$$

According to the Bayesian Network in the full graphical model as shown in Figure 4, the likelihood $Pr(D|\theta_L)$ and prior $Pr(\theta_L|\beta_z)$ of the left arm can be further decomposed as: $Pr(D|\theta_L) = Pr(D_{LH}|\theta_{LH})Pr(D_{LUA}|\theta_{LUA})Pr(D_{LFA}|\theta_{LFA})Pr(D_{LS}|\theta_{LS})$ and $Pr(\theta_L|\beta_z) = Pr(\theta_{LH}|\theta_{LFA},\beta_z)Pr(\theta_{LFA}|\theta_{LUA},\beta_z)$ $Pr(\theta_{LUA}|\theta_{LS},\beta_z)$.

In the same manner, the likelihood $Pr(D|\theta_R)$ and prior $Pr(\theta_R|\beta_z)$ of the right arm can also be decomposed.

3.3 Inference for Arm Pose and Action Category

The pipeline of action categorization based on arm pose modeling is illustrated in Figure 3. At the first stage, the topic variable z i.e. action category is assumed to be observed. According to the Bayesian rule, the posterior probability of arm pose is proportional to its joint probability:

$$Pr(w|D,\beta,z) \sim Pr(D,w,\beta,z). \qquad (3)$$

For every hypothesis of action category $z \in \{1,2,\cdots,K\}$, MAP is used for arm pose modeling as follows,

$$\hat{w}_z = argmax_w Pr(w|D,\beta,z) = argmax_w Pr(D,w,\beta,z). \quad (4)$$

There will be a set of results for arm pose modeling, $\hat{w}_z \in \{\hat{w}_1,\cdots,\hat{w}_K\}$ and the corresponding joint probabilities $Pr(D,\hat{w}_z,\beta,z) \in \{Pr(D,\hat{w}_1,\beta,z=1),\cdots,Pr(D,\hat{w}_K,\beta,z=K)\}$. A dynamic programming is used to infer the arm pose parameters, and it will be descripted in Section 4.3. Then, action categorization is based on the results of arm pose modeling.

At the second stage, the latent variable i.e. arm pose parameter w is assumed to be observed. For every hypothesis of arm pose parameter $\hat{w}_z$ and its corresponding probability $Pr(\hat{w}_z|D,\beta)$, the final action category $\hat{z}$ is given by

$$\hat{z} = argmax_z Pr(z|\hat{w}_z) Pr(\hat{w}_z|D,\beta), \qquad (5)$$

where $Pr(\hat{w}_z|D,\beta)=\frac{Pr(\hat{w}_z,D,\beta,z)}{\sum_{k=1}^{K} Pr(\hat{w}_z,D,\beta,z=k)}$ is the probability of the arm pose $\hat{w}_z$ to be the final arm pose modeling, and $Pr(z|\hat{w}_z)$ is the probability that the estimated arm pose $\hat{w}_z$ is classified to action category z. $Pr(z|\hat{w}_z)$ is trained by a soft-max model which will be descripted in Section 4.4. Then the related arm pose parameter $\hat{w}_{\hat{z}}$ is the final result of arm pose modeling for the current visual observation D.

4 ARM POSE MODELING AND ACTION VALIDATION

4.1 Prior Estimation

A Gaussian kernel based non-parametric distribution (Li and Yung, 2012) is derived for arm priors from a training data set. If the size of the training data is large enough, the non-parametric distribution estimate is suitable to represent the required distribution. However, in many applications, training data is sparse or is not easy to collect. Therefore, semi-parametric distribution estimation (Scarrott and MacDonald, 2012) is one of the methods used to estimate the required distribution. Generally, there are more observations in regions with a high density of data than in regions with low density of data. In the tails of a distribution where data are sparse, the non-parametric estimate performs poorly. In this case, the semi-parametric distribution estimate takes advantage of both the parametric estimate and non-parametric estimate. In the center of the distribution, a non-parametric estimate such as Gaussian kernel based estimate is used to estimate the cumulative density function (CDF). A parametric estimate such as a generalized Pareto distribution (GPD) is then employed for each tail.

The probability density function of variable x for the GPD with shape parameter k, scale parameter σ and threshold parameter μ, is

$$f(x|k,\sigma,\mu) = \begin{cases} \frac{1}{\sigma}(1+k\frac{x-\mu}{\sigma})^{-\frac{k+1}{k}}, & k \neq 0 \\ \frac{1}{\sigma}exp(-\frac{x-\mu}{\sigma}), & k = 0 \end{cases}. \qquad (6)$$

where $\mu < x$ when $k \geqslant 0$ or $\mu < x < -\frac{\sigma}{k}$ when $k < 0$. If $k = 0$ and $\mu = 0$, the GPD is equivalent to the exponential distribution. If $k > 0$ and $\mu = \frac{\sigma}{k}$, the GPD is equivalent to the Pareto distribution. The parameters of generalized Pareto can be estimated by the maximum likelihood estimation (Davison and Smith, 1990). Finally, the estimated semi-parametric distributions combined with uniform distributions are normalized and discretized as the priors of arm pose.

4.2 Likelihood Computation

The overall likelihood of arm parts comes from two types of evidence: the evidence from the lines and regions and the evidence from the pixel-based dense features.

4.2.1 Likelihood from Lines and Regions based Features

To derive the likelihood of an image patch containing an arm part, the boundary and foreground information for the upper arm and forearm, and skin color for the hand are selected to be the salient features. Since upper arm and forearm are more likely to be covered by sleeves of clothes, their color or texture information is unreliable. In this regard, boundary and foreground features are used for the upper arm and forearm instead, while color information is mainly used for the hand.

To evaluate the boundary, foreground and skin color features on a patch, two types of probabilistic templates are proposed (Li and Yung, 2012). The features are boundary, foreground and skin color mask. The probability boundary (*pb*) approach (Martin et al., 2004) is used to generate boundary feature f_{pb}, the foreground f_{fg} is extracted by the method (Wang and Yung, 2010), and the skin color mask f_{sc} is produced by the method (Conaire et al., 2007). Two probabilistic templates are proposed to calculate arm parts' likelihoods with the derived features. One template $b_{\theta\ldots}$ contains two Gaussian distributions on both sides for boundary features, and another template $g_{\theta\ldots}$ contains only Gaussian distributions in the middle for

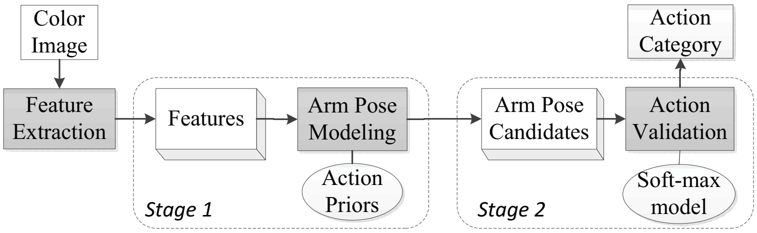


Figure 3: The pipeline for action categorization based on arm pose modeling.

foreground and skin mask, where $\theta_{\dots}$ is the parameter of an arm part. All the probabilistic templates are normalized.

4.2.2 Likelihood from Dense Features

To derive more reliable evidence of the appearance of arm parts, pixel based dense features are incorporated. In this proposed method, the combined feature descriptors used are local binary patterns (LBP), scale-invariant feature transform (SIFT), colour-SIFT and texton. To derive the confidence value between a pixel and a specific class label, the Joint Boosting algorithm (Torralba et al., 2004) is adopted which is an efficient approach to train multi-classifiers jointly by finding common features that can be shared across classes. The confidence value of class c and the combined feature x_i for pixel i is derived by a learned strong classifier in an additive model of the form $H(x_i,c) = \sum_{m=1}^{M} h(x_i,c)$, summing the classification confidence of M weak classifiers. Actually, it is the weight of the edge between a category label and the dense features of a pixel as shown in Figure 4.

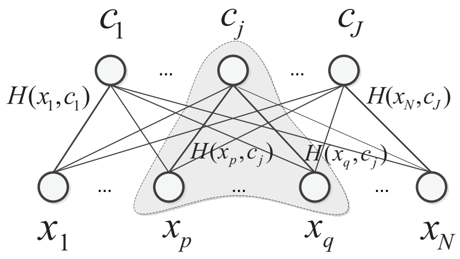


Figure 4: Bipartite graph of pixel based dense features and category labels, where $c_1,\cdots,c_j,\cdots,c_J$ is the set of labels with J categories, $x_1,\cdots,x_p,\cdots,x_N$ is the set of dense features for N pixels, and the edges between two sets are the confidence values H from the trained Joint Boosting classifiers. The shaded part is a sub bipartite graph with only one category label and the dense features of some pixels.

For a given parameter θ of a specific arm part c_j, if the involved dense features are represented by $X = \{x_p,\cdots,x_q\}$, the likelihood based on dense features can be calculated by

$$\begin{aligned} Pr(f_{ds}|\theta_{c_j}) = Pr(X|c_j) &= \frac{\sum_{x\in X} Pr(x,c_j)}{\sum_{x\in ds} Pr(x,c_j)} \\ &= \frac{\sum_{x\in X} exp(H(x,c_j))}{\sum_{x\in ds} exp(H(x,c_j))}, \end{aligned} \tag{7}$$

where $f_{ds} = \{x_1,x_2,\cdots,x_N\}$ is the dense features for all pixels.

4.2.3 The Overall Likelihood

The likelihoods of upper arm (UA), forearm (FA), and hand (H) are given as below:

$$\begin{aligned} Pr(D|\theta_{UA}) &= Pr(f_{pb},f_{fg},f_{ds}|\theta_{UA}) \\ &= Pr(f_{pb}|\theta_{UA})Pr(f_{fg}|\theta_{UA})Pr(f_{ds}|\theta_{UA}), \end{aligned} \tag{8}$$

$$\begin{aligned} Pr(D|\theta_{FA}) &= Pr(f_{pb},f_{fg},f_{ds}|\theta_{FA}) \\ &= Pr(f_{pb}|\theta_{FA})Pr(f_{fg}|\theta_{FA})Pr(f_{ds}|\theta_{FA}), \end{aligned} \tag{9}$$

$$Pr(D|\theta_{H}) = Pr(f_{sc}|\theta_{H})Pr(f_{ds}|\theta_{H}). \tag{10}$$

The upper arm UA can be the left upper arm LUA or the right upper arm RUA, and the forearm FA and hand H also can be one of the two arms. These items are incorporated in the expansion of Eqt.2 to calculate the likelihoods for the left and right arms.

4.3 Inference for Arm Pose

When an action hypothesis is given, its pose modeling can be derived by MAP based on the extracted features and the corresponding priors. Dynamic programming (Felzenszwalb and Zabih, 2011), one of message passing methods, is efficient enough to solve the MAP problem. As shown in Figure 5, there are three main layers for the lattice of one arm pose estimation. The three layers from left to right represent the parameter states of UA, FA and H. It also has a start node P_S which represents the shoulder location parameter, and an end node P_E. Similarly in every layer, nodes represent all parameter states that the corresponding arm part may hold. For example, θ_{UA}^{m} represents the m^{th} parameter state for the upper arm which has M different parameter states in all. Besides, the node will assign a score S_{UA}^{m} during inference. There are directed edges between nodes of

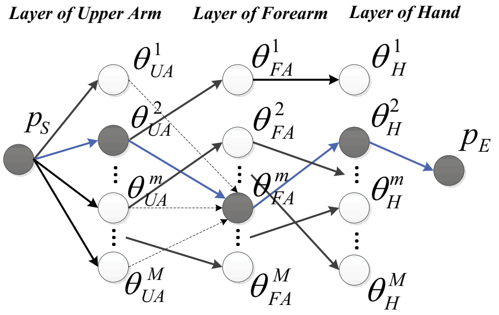


Figure 5: Lattice structure for dynamic programming.

adjacent layers and their weights represent the conditional probability between arm parts. Dynamic programming starts from the start node P_S, and along the layer direction to the end node P_E. The detailed procedure is as follows.

The score of the start node is initialized as $S_{P_s} = 0$, then the score of the m^{th} node in the second layer, i.e. upper arm layer, is

$$S^m_{UA} = S_{P_s} + logPr(\theta^m_{UA}|P_s) + logPr(D|\theta^m_{UA}). \quad (11)$$

After computing the scores of nodes in the layer of upper arm, the score of the mth node in the third layer, i.e. forearm layer, is

$$S^m_{FA} = max_n \; S^n_{UA} + logPr(\theta^m_{FA}|\theta^n_{UA}) + logPr(D|\theta^m_{FA}). \quad (12)$$

Then, in the same manner, the score of the mth node in the layer of hand, i.e. the third layer, is

$$S^m_H = max_n \; S^n_{FA} + logPr(\theta^m_H|\theta^n_{FA}) + logPr(D|\theta^m_H). \quad (13)$$

Finally, the score of the end node P_E is

$$S_{P_E} = max_n \; S^n_H. \quad (14)$$

During inference, all the nodes record their previous nodes which contribute to the maximum. So, the route can be retraced from the end node based on the records of the previous nodes. Then the nodes in the corresponding route with the maximum posterior are the estimated parameters $\hat{\theta}_{LA}$ or $\hat{\theta}_{RA}$ for the MAP solution. This procedure makes MAP inference possible and efficient.

4.4 Action Hypothesis Validation

Soft-max regression (Duan et al., 2003) is an efficient approach for multi-class classification and generalizes logistic model where the class label can take on more than one possible value. In our problem, the estimated parameter of arm pose $\hat{w}_z$ is the input to the soft-max regression model, and it produces the hypothesis of the probabilities $Pr(y = k|\hat{w}_z)$ for the k^{th} action category. The probability $Pr(y = z|\hat{w}_z)$ is used for action category validation.

Training soft-max regression model is a supervised procedure. The training set is $\{w_k\}$ which contains the samples of arm pose and their corresponding action category $k \in \{1, 2, \cdots, K\}$. The i^{th} pose sample of the action category k is represented by w^i_k and its hypothesis is

$$\begin{aligned} h_\vartheta(w^i_k) &= [Pr(y = 1|w^i_k; \vartheta), \cdots, Pr(y = K|w^i_k; \vartheta)]^T \\ &= \frac{[exp(\vartheta_1^T w^i_k), \cdots, exp(\vartheta_K^T w^i_k)]^T}{\sum_{j=1}^K exp(\vartheta_j^T w^i_k)}, \end{aligned} \quad (15)$$

where $\vartheta = [\vartheta_1, \cdots, \vartheta_K]^T$ is the soft-max model's parameter, and $\sum_{j=1}^K exp(\vartheta_j^T w^i_k)$ is used for normalization. The model parameter ϑ can be optimized by gradient descent using training samples and their labels. For an estimated arm pose w, $h_\vartheta(w)$ gives the probabilities that arm pose w belongs to every action category.

After the first stage of arm pose modeling, for each action category z, there is a corresponding estimated arm pose $\hat{w}_z$. And $h_\vartheta(\hat{w}_z)$ can be derived based on the trained soft-max model. Then $Pr(z|\hat{w}_z)$ in Eqt.5 for action category validation is

$$Pr(z|\hat{w}_z) = h^z_\vartheta(\hat{w}_z) = \frac{exp(\vartheta_z^T \hat{w}_z)}{\sum_{i=1}^K exp(\vartheta_z^T \hat{w}_z)}. \quad (16)$$

5 EXPERIMENT AND RESULT

The dataset used for evaluating the proposed method is the HumanEva-I (Sigal and Black, 2006). It consists of mainly frontal images of three actions: Walking, Jogging and Waving (a subset of Gesture); and each image is annotated with positions of shoulder, elbow, wrist and hand endpoint for both arms. There are four subjects in the dataset and their appearances vary significantly in style, type, and color of clothing. The number of frontal images for one action of a specific subject is about 100, of which half of them are selected for training and the other half for testing. To train the potentials of pixel based labeling, the boundaries of upper body parts are needed. This includes upper arms, low arms, hands and torso for both arms, and they are approximated by rectangles connected by the annotated joints, and the head is approximated by a circle. The remaining region is annotated as background.

5.1 Arm Pose Modeling Result

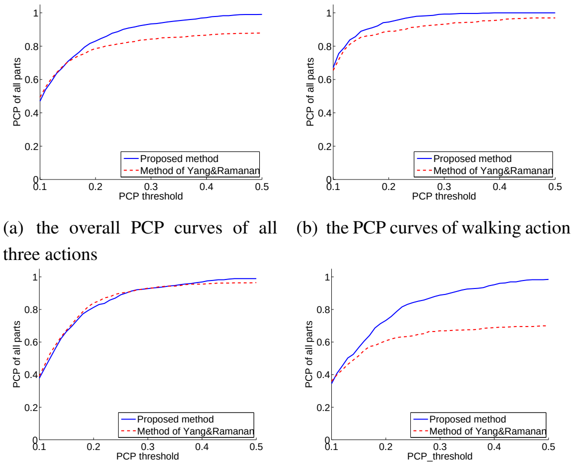
Percentage of Correctly estimated body Parts (PCP) (Ferrari et al., 2008) is one of the most popular measures for 2D pose estimation which is adopted for arm

Table 1: The average PCP for three actions and each arm part of both arms for proposed method and Yang&Ramanan's method.

Average PCP	Actions			Arm Parts				Overall
	Walking	Jogging	Waving	LUA	RUA	LFA	RFA	
Proposed	95.49%	85.85%	81.8%	95.55%	95.42%	77.24%	82.27%	87.62%
Yang&Ramanan	90.97%	85.94%	60.75%	93.83%	85.04%	73.91%	68.85%	80.4%

pose evaluation. To evaluate the performance of our approach, the approach of mixtures of parts (Yang and Ramanan, 2011) as one of the state-of-art methods is selected as reference. It uses mixtures of parts based on the histogram of oriented gradients (HOG) descriptor.

According to Figure 6(a), the overall PCP curves show our method has better performance than the method of Yang and Ramanan on the testing images and improves about 7.21% average PCP. In details for different actions, proposed method improves 4.5% and 19.05% on walking and waving action. As illustrated in Table 1, our method gains 95.49%, 85.85%, and 81.8% PCP for the actions walking, jogging and waving respectively. Generally, jogging and walking actions have smaller variations in space than waving actions. This proposed adopted the learnt prior to capture the possible variation for all actions.



(a) the overall PCP curves of all three actions

(b) the PCP curves of walking action

(c) the PCP curves of jogging action

(d) the PCP curves of waving action

Figure 6: The overall PCP curves for all three actions and the individual PCP curves for every action in which the blue curve and the red dotted curve are the results of proposed method and Yang & Ramanan's method respectively.

As shown in Table 1, two methods all have better performance for upper arm than forearm of both arms. Proposed method has 95.48% and 79.76% average PCP for upper arm and forearm, while method of Yang and Ramanan has 89.44% and 71.38% average PCP respectively.

5.2 Action Categorization Result

After the arm poses modeling for every possible action, the final result of action category is the action with maximum probability of arm pose modeling for the current visual observation. Since in this research, we mainly focus on the arm pose estimation for action categorization. The images of walking, waving and jogging from the HumanEva-I dataset are tested. Table 2 is the confusion matrix for this three action categorization. The average recognition rate is about 96.69%, and it has best performance for walking action. Figures 7, 8 and 9 illustrate some arm pose modeling for the actions waving, jogging and walking respectively. Recently, the approach of Bag of Poses (BoP) (Gong et al., 2013) is used for action recognition by SVM classifiers on this dataset. It has the recognition rates 94.6%, 91.9%, and 91.8% for Walk-

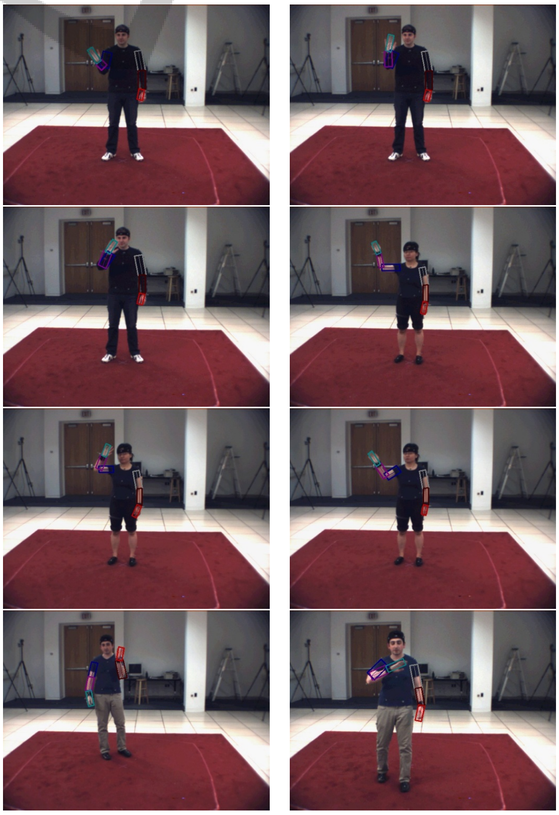


Figure 7: Some examples of arm pose modeling for waving.

Table 2: The confusion matrix for action categorization of the actions walking, waving, and jogging of the HumanEva-I dataset.

Acc.	Walking	Waving	Jogging
Walking	99.35%	0.65%	0
Waving	4.28%	94.55%	1.17%
Jogging	3.83%	0	96.17%

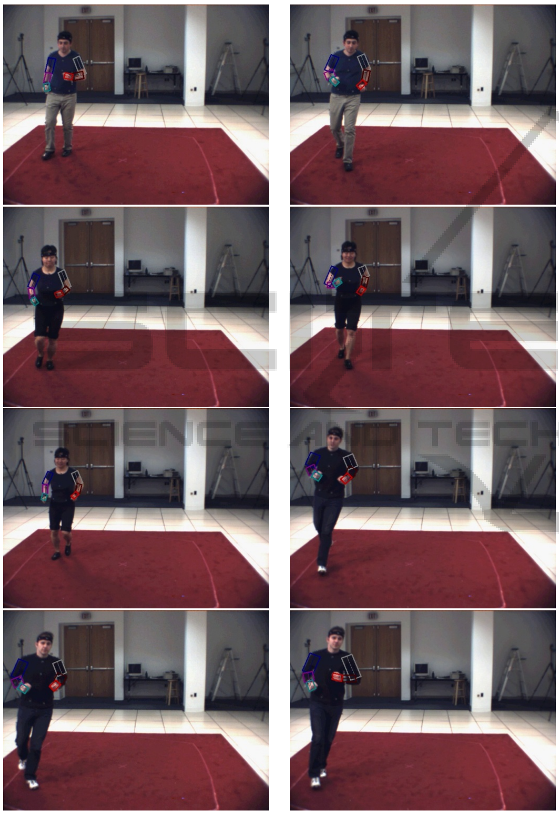


Figure 8: Some examples of arm pose modeling for jogging.

ing, Gesture (mainly waving action) and Jogging respectively.

6 CONCLUSIONS

This paper proposed a novel method to implement the categorization of actions such as waving, walking and jogging, with the help of arm pose modeling. Unlike many existing methods, we treat pose modeling and action recognition interdependently. Proposed method explored the relationship between arm pose modeling and action categorization, as well as multiple visual features and priors for arm pose modeling. We utilized a graphical model to descript relationship between arm pose and action category, and the inherent dependency between arm parts. Some new methods of prior distribution estimation, likelihood calculation, and the inference for arm pose and action category were illustrated. This method was evaluated on the videos of walking, waving and jogging from the HumanEva-I dataset. It improved 7.21% average PCP over the method of Yang and Ramanan for arm pose modeling, and achieved 96.69% average action categorization rate. The result approved that our arm pose modeling is useful for action categorization, and the priors of action category can benefit arm pose modeling conversely. For future research, if there are enough training samples for prior distribution estimation, the imbalance problem will be alleviated for action categorization. Moreover, this paper only shows its efficiency on three actions and the arm pose modeling. More complex actions or modeling for the whole body pose from different viewpoint will be considered.

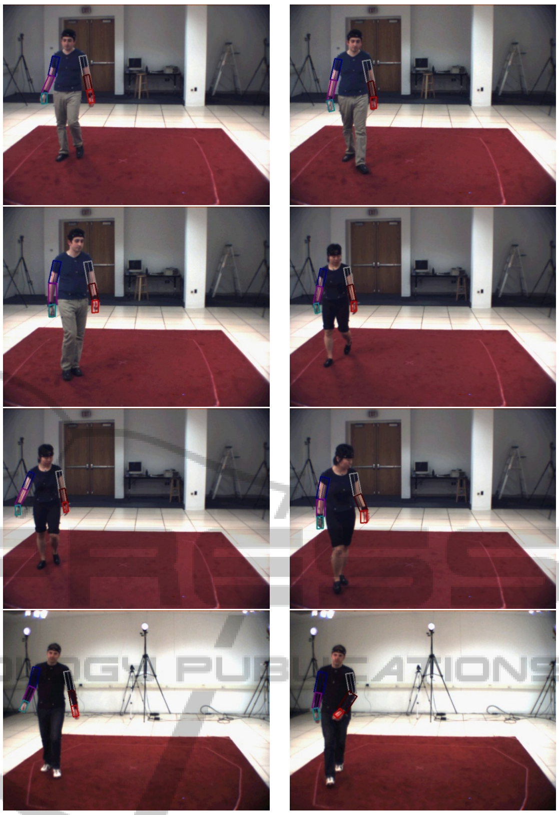


Figure 9: Some examples of arm pose modeling for walking.

ACKNOWLEDGEMENTS

This research is supported by the Postgraduate Studentship of the University of Hong Kong.

REFERENCES

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *ICCV*, volume 2, pages 1395–1402. IEEE.

Conaire, C. O., O'Connor, N. E., and Smeaton, A. F. (2007). Detector adaptation by maximising agreement between independent data sources. In *CVPR*, pages 1–6. IEEE.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 393–442.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72. IEEE.

Duan, K., Keerthi, S. S., Chu, W., Shevade, S. K., and Poo, A. N. (2003). Multi-category classification by softmax combination of binary classifiers. In *Multiple Classifier Systems*, pages 125–134. Springer.

Elgammal, A., Shet, V., Yacoob, Y., and Davis, L. S. (2003). Learning dynamics for exemplar-based gesture recognition. In *CVPR*, volume 1, pages I–571. IEEE.

Fathi, A. and Mori, G. (2008). Action recognition by learning mid-level motion features. In *CVPR*, pages 1–8. IEEE.

Felzenszwalb, P. F. and Zabih, R. (2011). Dynamic programming and graph algorithms in computer vision. *PAMI*, 33(4):721–740.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *CVPR*, pages 1–8. IEEE.

Gong, W. et al. (2013). *3D Motion Data aided Human Action Recognition and Pose Estimation*. PhD thesis, Universitat Autònoma de Barcelona.

Laptev, I. (2005). On space-time interest points. *IJCV*, 64(2-3):107–123.

Li, C. and Yung, N. (2012). Arm pose modeling for visual surveillance. In *IPCV*, pages 340–347.

Martin, D. R., Fowlkes, C. C., and Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26(5):530–549.

Moeslund, T. B., Hilton, A., Krüger, V., and Sigal, L. (2011). *Visual analysis of humans: looking at people*. Springer.

Natarajan, P. and Nevatia, R. (2012). Hierarchical multichannel hidden semi markov graphical models for activity recognition. *CVIU*.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318.

Rodriguez, M., Ahmed, J., and Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8.

Sadanand, S. and Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *CVPR*, pages 1234–1241. IEEE.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In *ICPR*, volume 3, pages 32–36. IEEE.

Sigal, L. and Black, M. J. (2006). Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 120.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, volume 2, pages II–762. IEEE.

Wang, L. and Yung, N. H. (2010). Extraction of moving objects from their background based on multiple adaptive thresholds and boundary evaluation. *ITS*, 11(1):40–51.

Xu, R., Agarwal, P., Kumar, S., Krovi, V. N., and Corso, J. J. (2012). Combining skeletal pose with local motion for human activity recognition. In *Articulated Motion and Deformable Objects*, pages 114–123. Springer.

Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In *CVPR*, pages 379–385. IEEE.

Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392. IEEE.

Yao, A., Gall, J., Fanelli, G., and Van Gool, L. (2011). Does human action recognition benefit from pose estimation?". In *BMVC*, pages 67.1–67.11.