A novel human behaviour information coding method based on eye-tracking technology

Tao Hu, Jian Lv, Qingsheng Xie*, Hui Sun, Qingni Yuan

Key Laboratory of Advanced Manufacturing Technology,
Ministry of Education, Guizhou University, Guiyang, 550025, China

60630571@qq.com

ABSTRACT. *The computer recognition and coding of human motion has great application potentials in intelligent human-computer interaction (HCI), virtual reality and computer vision. The advancement of these fields hinges on the development of computers capable of processing information like the human brain. However, there is little report on computer coding of human motion based on the information processing features of human vision. Therefore, this paper attempts to develop a computer vision information coding model for recognizing human actions in virtual situations.*

Moreover, human motion was defined in terms of concept, vision and perception, and a visual cognition experiment was performed to explore the influence of different contexts on motion information recognition. The main results are as follows: the fixation duration is nonlinearly correlated with context; interaction occurs between different contexts in the target motion recognition process; the recognition efficiency was higher at the appearance of the target motion with the context (of similar motions) than at that of the target motion with the mixed context. Based on the experimental data, the author designed strategies for visual information feature coding, information classification and information processing. These strategies enable computer vision to recognize and process information under complex background in an accurate and efficient manner. The research findings lay the basis for intelligent HCI and information visualization in virtual environment.

RÉSUMÉ. *La reconnaissance et le codage informatiques du mouvement humain ont un potentiel d'application considérable dans les domaines de l'interaction intelligente homme-machine (IHM), de la réalité virtuelle et de la vision par ordinateur. L'avancement de ces domaines dépend du développement d'ordinateurs capables de traiter des informations telles que le cerveau humain. Cependant, il existe peu de rapports sur le codage informatique du mouvement humain basé sur les caractéristiques de traitement de l'information de la vision humaine. Par conséquent, cet article tente de développer un modèle de codage des informations de vision informatique permettant de reconnaître les actions humaines dans des situations virtuelles.*

De plus, le mouvement humain a été défini en termes de concept, de vision et de perception, et une expérience de cognition visuelle a été réalisée pour explorer l'influence de différents contextes sur la reconnaissance des informations de mouvement. Les principaux résultats sont les suivants: la durée de la fixation est en corrélation non linéaire avec le contexte; l'interaction

se produit entre différents contextes dans le processus de reconnaissance de mouvement cible; l'efficacité de la reconnaissance était plus élevée à l'apparition du mouvement cible au contexte unique (de mouvements similaires) qu'à celle du mouvement cible au contexte mixte. Basé sur les données expérimentales, l'auteur a conçu des stratégies de sélection de caractéristique des informations visuelles, de classification et de traitement de l'information. Ces stratégies permettent à la vision par ordinateur de reconnaître et de traiter des informations de manière précise et efficace dans des contextes complexes. Les résultats de la recherche jettent les bases de l'interaction intelligente homme-machine et de la visualisation d'informations dans un environnement virtuel.

KEYWORDS: *information identification, information coding, motion capture, fixation duration, virtual reality*

MOTS-CLÉS: *identification de l'information, codage de l'information, capture de mouvement, durée de fixation, réalité virtuelle*

DOI:10.3166/TS.34.153-173

1. Introduction

The recognition and coding of human motion signals mark the future of technologies like intelligent human-computer interaction (HCI) and computer vision, which are of great application potentials in industrial design, intelligent robot and intelligent manufacturing. With the development of virtual reality technology, the recognition of human behaviours in virtual scenes has become a key focal point in HCI research.

The past few decades have seen the emergence of various motion recognition approaches. These approaches are grounded on different bases, such as space-time volume, space-time, facet, graph graph-based model, critical pose, kernel and deep learning (Chen, 2016). Despite the research progress, it is still difficult to recognize human behaviours accurately from motion sequences.

In recent years, more and more scholars have attempted to identify motions based on skeleton information, e.g. the motion sequence diagrams prepared from human skeleton features, which are extracted by human motion capture systems. Some of these systems are based on the depth map, some on bone joints, some on joint group, and some others on joint dynamics. The space information is often captured by Fourier temporal pyramid (FTP), while the temporal information is usually simulated by hidden Markov models (HMMs) (Shotton *et al*., 2011). Nevertheless, these manually extracted features are shallow, dependent on data, or unlearned in an end-to-end manner.

Some of the existing studies on skeleton motion recognition identify motion information and user behaviour through image and motion segmentation. For instance, Reference (Chen and Chen, 2011) acquires the image boundaries via wavelet analysis and obtains the accurate boundaries of the human body by mathematical morphology. Reference (Zhang *et al*., 2014) determines the motion boundaries in light of the eigenvectors on three locations of the human body. Nonetheless, the segmented image quality is too poor to realize effective recognition of multi-motion sequences and

motion scenes. This calls for further improvement of the recognition accuracy of skeletal motion sequence, which can be viewed as a set of independent frames with certain smoothness in time domain, a subspace of poses or pose features, or the output of the neural network encoder.

In light of the above, this paper puts forward a multi-context feature coding method for behavioural identification and motion information classification, and applies it to construct a behavioural information coding model in virtual scenes. The recognition efficiency of the proposed method was tested through a cognitive experiment on a target motion in a similar context (containing similar motions) and a mixed context (containing significantly different motions). During the experiment, the participants were presented with three sets of target motions and four contexts, and asked to determine the presence of the target action and find its local context. It is assumed that participants could identify the target motion more efficiently in the similar context than in the mixed context, because the combination of similar motions forms a unified perception when the human brain recognizes the motions. In addition, a nonlinear equation was adopted to depict the relationship between different contexts and the fixation duration of the participants, aiming to examine the experimental effects of the proposed method in a comprehensive manner.

The remainder of this paper is organized as follows: Section 2 reviews the previous research of human motion recognition, interpretation and classification; Section 3 introduces the information coding model, the information storage mechanism and the cognitive experiment; Section 4 presents the experimental results and analyses the data; Section 5 discusses the coding strategies; Section 6 wraps up this paper with meaningful conclusions.

2. Literature review

2.1. Human motion recognition

As a key issue in computer vision, human motion recognition is widely used in video monitoring, the human-machine interface, robotics etc., laying the basis for information coding in this research. Below is a brief review of recent studies on this issue.

Using depth image recognition, Reference (Barbosa *et al*., 2012) acquires the facial, appearance and bone information of a user, compares the information with the current body information of the user, and identifies the user through confidence calculation and motion capture. Reference (Zhu and Kikuo, 2010) introduces depth information to improve the accuracy and robustness of traditional Bayesian motion tracking in the tracing of human body motions. To promote traffic safety, Reference constructs a novel driving behaviour recognition system based on a specific physical model and motion sensory data.

Reference combines phone acceleration sensor and deep convolutional network (CNN) into a behaviour identification and classification model, which consists of an

input layer, two convolutional layers, two pool layers, a fully-connected layer and an output layer; the model classifies motions against the data features extracted by sliding window method. References (Cheng *et al*., 2013) rely on the latest positioning techniques and phone sensors to capture human motions in natural environment, and investigate human behaviours using the captured motions.

Reference (Valstar, 2015) sets up a continuous dynamic system framework that supports single motion simulation and two-action conversion, and uses the system to identify the type and boundaries of the target motion. References (Zhang *et al*., 2015) identify the static features of the target motion, such as size, colour, edge, contour, shape and depth. To capture motion features, Reference (Hemati and Mirzakuchaki, 2014) employs the Fiedler embedding method to nest a spin image and a local cube in the same space.

Reference (Wang *et al*., 2014) recognizes human motion considering the correlation difference between the joints under a joint group.

Reference (Zanfir *et al*., 2013) expresses the time-varying joint positions with a continuous and differentiable function (a.k.a. skeletal motion sequence); the function is constructed through the following steps: computing the shape context features using the isometric points subsampled from body segments, depicting the human skeleton as a sequence of time-varying features (e.g. position, tangent and shape), splitting the time sequence into several sub-sequences, each of which is modelled by a linear dynamic system, and using the parameters to create a dynamic skeletal motion sequence.

Reference (Slama *et al*., 2015) represents the joint trajectories of each motion sequence as a linear dynamic system, and expresses the system as the observable matrix of embedded model parameters by the autoregressive-moving-average model. Reference (Lehrmann *et al*., 2014) applies the dynamic forest model and several autoregressive tree models for motion recognition; each node of the probabilistic autoregressive tree stores a multivariate normal distribution in the form of a fixed covariance matrix, and calculates the result based on the Gaussian distribution estimated by the forest.

2.2. Human motion interpretation and classification

Many novel methods have been developed to interpret and classify human motions through deep learning. The following are some of the representative studies on human motion interpretation and classification with the aid of deep neural networks.

Reference (Ji *et al*., 2013) proposes several ways to analyse the motion video: treating it as a series of static images, dividing the motion sequence into continuous sub-sequences with smooth transitions, inputting it as volume to the CNN extended to the third time dimension. Reference (Bilen *et al*., 2016) designs a long-term pool under the neural network, which compresses the video sequence into a single dynamic image and sorts the image according to the time sequence of the video frames. Reference (Fr and Landragin, 2006) suggests that human behaviours can be accurately

identified by the dynamic light information of joints. References develop a way to identify motions in light of the historical trajectories of key points.

The motion recognition has advanced greatly thanks to techniques involving robust feature extractors and related descriptors, including scale-invariant feature transformation (SIFT), the histogram of oriented gradients (HOG), local binary pattern (LBP) and revised pyramid histogram visual word (PHOW). These techniques have been adopted to analyse the motion sequence in motion recognition. For example, Reference (Loncomilla *et al*., 2016) adopts 3D SIFT for motion recognition using temporal and spatial properties. Reference (Scovanner *et al*., 2003) extends the LBP into volume LBP (VLBP) to integrate motion with appearance attributes, and only considers the commonness of three separate layers while applying the method to identify dynamic features, aiming to reduce the computing load and enhance the scalability.

Reference (Almaev and Valstar, 2013) proposes a method to recognize the general and specific human behaviours in motion sequences: first, the body postures and body part positions are assessed in the video sequence; then, image patches of different sizes (three from RGB and three from OF) are used for block classification; finally, each image patch is processed according to the corresponding CNN information on appearance and motion.

Reference puts forward the strategies of sequence code maps (SCMs) and relative location probabilities (RLPs), both of which are efficient ways to describe the relationship between motion features. References (Matikainen *et al*., 2010) present a behavioural feature depiction method that can identify motions and behaviours.

The identification of human motions, continuous or discrete, is comparable to data classification. With the aid of template matching classification, References (Maji *et al.*, 2011) characterize the image sequence by kinetic energy map and a motion history map, and compute the distance between two templates by the Mahalanobis distance equation (Althloothi *et al*., 2014). References (Bobick and Davis, 2001; Jiang *et al*., 2012) adopt the adjacent classification method, a.k.a. the template allocation method to calculate the descriptor distance between images in the observed sequence and the training sequence.

3. Methodology

3.1. Information coding model

Our information model was constructed by coding the behaviour information in the virtual context, following the agent modelling theory. The model structure is represented as image, processing depth (PD), context, connection weight and route. These factors are detailed as follows.

3.1.1. Image

The image refers to the memory or perceptual image in cognitive activities. It is an instrument to represent perceptual information. Here, an image is defined as the set of user behaviours (i.e. actions or motions) in a virtual scene, which can be coded, stored and extracted:

$$Iage = \left\{ \left(Conte_1,\ Act_1 \right), \left(Conte_2,\ Act_2 \right) \cdots \left(Conte_i,\ Act_i \right) \right\}.$$

where *Iage* is the image set; Act_i is the user action; $Conte_i$ is the virtual scene of the user behaviour.

3.1.2. PD

The PD is an attribute of the coding operation for independent item information correlated with multiple external factors. Here, this term is defined as the degree of enhancement of the connection between the image information inputted to the coding channel and the outside world. The PD was adopted to represent the connection between information images and the effect of information extraction and processing.

$$Dop_i = \left\{ Iage \middle| Iage\left(Conte_j \right) \equiv Conte_i,\ \ Conte_j \to Conte_i \right\}.$$

where Dop_i is the PD; *Iage* is an image; $Conte_j$ and $Conte_i$ are the contexts of the current information input and the information output, respectively.

3.1.3. Context

The context, related to the real-time state of the cognition system, describes the environment of the explicit/implicit cognition target. Here, this term is defined as the space for HCI activities, which has a similar state as the user.

$$Conte_{i_k} = \left\{ Conte_{i_k} \middle| Conte_{i_k}\left(Act_i \right) \quad i \in [1,\ i],\ k \in [1,\ p_i] \right\}.$$

where $Conte_{i_k}$ is the current context information; i is the serial number of environmental factors; p_i is the current amount of context information; Act_i is the current behaviour.

3.1.4. Connection weight

As a concept of neuroscience, the connection weight measures how closely two neurons on different layers are correlated when the information is transmitted between them. Here, the connection weight illustrates how much an information image is affected by that on the superior layer during information transmission between the two layers. The value of connection weight varies with the learning intensity.

$$\Delta w_{ij} = \varepsilon\ E\ f\left(I \right)$$

where Δw_{ij} is the change of weight; ε is the learning rate; E is the error; $f(I)$ represents is the input image information.

3.1.5. Route

Also known as the information transmission channel, the route describes the relationship between information images.

$$Route = \left\{ R_j,\ R_i,\ w_{ij},\ f\left(I_j\right) \right\}.$$

where $Route$ is the information transmission channel; R_j is the input information; R_i is the output information; w_{ij} is the connection weight of the information; $f(I_j)$ is the information to be transmitted. A route comes into being when w_{ij} surpasses a pre-set threshold.

3.2. Information storage mechanism

3.2.1. Short-term storage

Short-term storage, a psychological term, is a limited capacity system that can simultaneously store, manage and process the temporary behaviour information. In this paper, the expected information is determined according to user behaviours acquired in the current situation. This information can be classified and coded in three categories, namely, concept, vision and perception, and saved in the short-term storage as information images. In this area, the information is encoded without further processing. The short-term storage is illustrated in Figure 1 below.

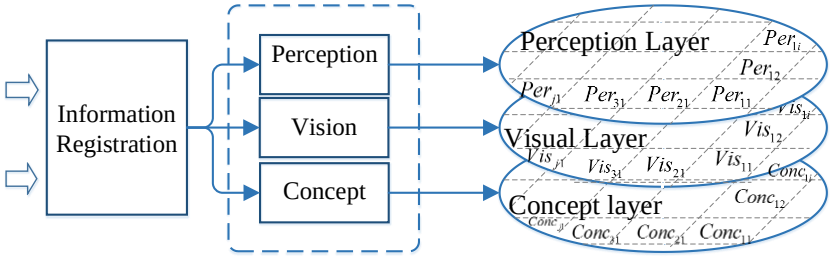

Figure 1. Short-term storage

3.2.2. Long-term storage

Long-term storage, also a psychological term, is a continuous working system that can record the constant changes of information in cognitive processing. The procedures and mechanism needed to complete the current tasks are embedded in long-term storage, making it possible to rapidly extract the user behaviours and guide

the user psychology in the long run. The long-term storage can simplify the procedures of short-term storage in treating complex cognitive activities.

3.3. Participants

Thirty students, 16 males and 14 females, were selected to participate in our experiment. All participants are computer literate and surpass 1.0 in naked or corrected visual acuity. None of them have eye problems like colour blindness.

3.4. Experimental design

3.4.1. Variables

An eye motion experiment was performed to determine the eye motion trajectories for further evaluation. The motion data were collected by motion capture sensors. Several independent variables were designed, including four contexts and the gender of the participants. The dependent variables, i.e. number of fixations, fixation duration and the participant's judgement performance of target motion, were used for the participants to judge the eye motions. The details of these variables are listed below:

(1) Independent variables:

-Four contexts: similar motion + target (*A*), similar motion (*B*), mixed motion + target (*C*) and mixed motion (D);

-Gender of the participants (*Gender*).

(2) Dependent variables:

The dependent variables include fixation duration (T), number of fixation (N), and the participant's judgement performance of target motion (*Judge*). Among them, the *Judge* was evaluated by the experimenter against the actual situation.

3.4.2. Target images

As shown in Figure 2, the target images were classified into three groups, each of which consists of 4 motions. In each group, 3 motions are similar while 1 motion is very different. The different motion is the target motion.

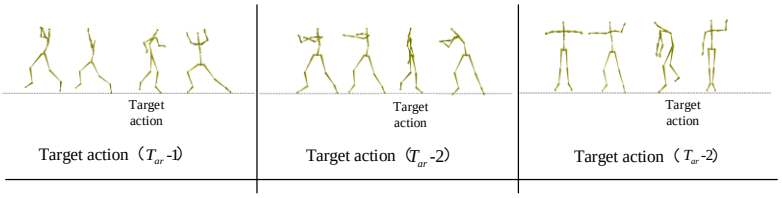

Figure 2. Target groups

3.4.3. Sample images

The sample images, denoted as *situation*, are composed of different contexts. These images were divided into three groups: Group 1: *A+B*; Group 2: *A+C*; Group 3: *B+D*. In each group, two areas of interest (AOIs) were defined, and denoted as area I and area II (Figure 3)

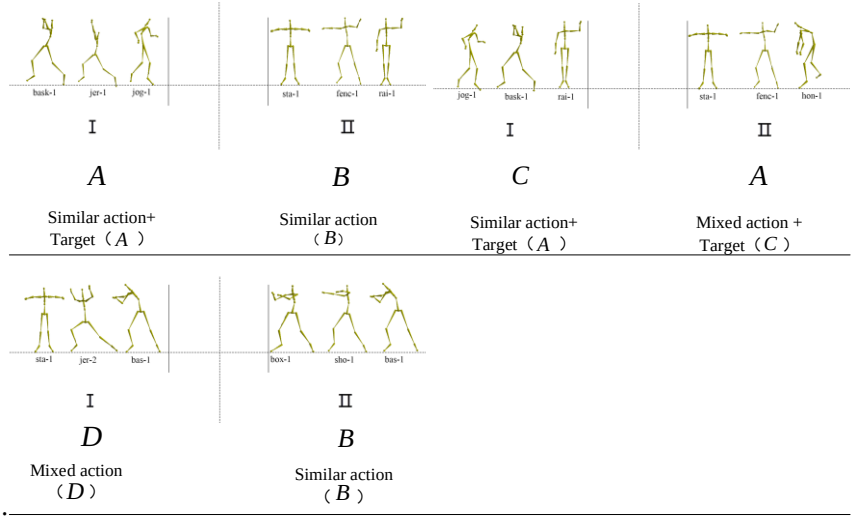

Figure 3. Sample groups

3.4.4. Experimental instruments

The experiment was carried out using an Eye So Ec60 (Braincraft Technology Co., Ltd.; sampling rate: 60Hz; sampling accuracy: 0.5°; number of calibration points: 9), a computer display (resolution: 1,028*1,024; refresh rate: 75Hz), and SPSS software.

The participants' eye motion trajectories, fixation duration (*T*) and the number of fixations were recorded, before determining the motion recognition strategy. The data on these parameters were all recorded by the Eye So, and the participants' judgement of the target motion was evaluated by the author.

3.4.5. Experimental procedure

First, three groups of target images were presented at an interval of 5s to the participants, followed by the three groups of sample images. After each group of sample images disappeared, each participant was asked to determine if the target motion appeared in the two AOIs.

Next, three groups of target images were presented at an interval of 5s to the participants, followed by the sample images. Each participant had to look for the target motion in the sample images, and press the space button on the keyboard if he/she

found the motion. After the sample images disappeared, the participant judged if the target motion was still present. After the judgement, the experimental task of these groups of target images was completed.

The above procedure was repeated until all sample images had been displayed to the participants.

4. Experimental results

4.1. Data analysis

The Box's M test results on the independent variables, i.e. gender and context, are listed in Table 1. It can be seen from the results $F=1.294$, $p=0.096$ and $p<0.05$ indicate that the covariance matrices of the two variables are equal.

Table 1. Box's M test results on the independent variables

Box's M	58.657
F Test	1.294
df 1	42
df 2	18628.955
Sig.	0.096

Table 2. Multivariate analysis of variance test results on the independent variables

Effect item		Value	F	Hypothesis df	Error df	Sig.
Intercept	Wilks' Lambda(λ)	0.16	3257.535b	3.000	158.000	0.000
Sex	Wilks' Lambda(λ)	0.909	5.286b	3.000	158.000	0.002
Context	Wilks' Lambda(λ)	0.713	6.369	9.000	384.681	0.000
Sex * context	Wilks' Lambda(λ)	0.973	0.478	9.000	384.681	0.890

Design: Intercept + sex + context + sex * context

Exact statistics

The static is the upper bound on F that yields a lower bound on the significance level.

Table 2 presents the multivariate analysis of variance test results on how significant the impacts from gender and context on dependent variables. The results on gender were Wilks' lamdba (λ) = 0.909, p=0.002 and p<0.05 and those on context were Wilks' (λ) = 0.713, p=0.000 and p<0.05, indicating that both variables have highly significant impacts on the dependent variables.

Table 3. Between-subjects effects test results on the independent variables

Source	Dependent variable	Type III sum of squares	df	Mean square	F	Sig.
Corrected Mode	Judge	4.265a	7	0.609	2.919	0.007
	Number of Eye Fixations	119.296b	7	17.042	4.358	0.000
	Fixation Duration	29506178.369c	7	4215168.338	9.212	0.000
Intercept	Judge	1617.787	1	1617.787	7750.843	0.000
	Number of Eye Fixations	4082.155	1	4082.155	1043.866	0.000
	Fixation Duration	410221243.813	1	410221243.813	896.511	0.000
Sex	Judge	2.382	1	2.382	11.412	0.001
	Number of Eye Fixations	21.584	1	21.584	5.519	0.020
	Fixation Duration	2525803.337	1	2525803.337	5.520	0.020
Context	Judge	0.967	3	0.322	1.544	0.205
	Number of Eye Fixations	94.949	3	31.650	8.093	0.000
	Fixation Duration	25372505.063	3	8457501.688	18.483	0.000
Error	Judge	33.396	160	0.209		
	Number of Eye Fixations	625.698	160	3.911		
	Fixation Duration	73212023.625	160	457575.148		
Sum	Judge	1911.000	168			
	Number of Eye Fixations	5365.000	168			
	Fixation Duration	572098803.000	168			
Corrected Total	Judge	37.661	167			
	Number of Eye Fixations	744.994	167			
	Fixation Duration	102718201.994	167			

R Squared = 0.113 (Adjusted R Squared = 0.074).

R Squared = 0.160 (Adjusted R Squared = 0.123).

R Squared = 0.287 (Adjusted R Squared = 0.256).

Table 3 shows the test results of between-subject's effects. The *p* values of gender relative to judgement performance, fixation duration and number of fixations were 0.001, 0.020 and 0.020, respectively, all of which are below 0.05. Hence, gender has highly significant impacts of gender on the dependent variables. The *p* values of context relative to judgement performance, fixation duration and number of fixations were 0.205, 0.000 and 0.000, respectively. Compared with the threshold of 0.05, it is learned that context has highly significant impacts on fixation duration and number of fixations, but does not significantly affect judgement performance.

4.2. Variance analysis

The multivariate analysis of variance suggests that the independent variables (gender and context) have highly significant impacts on the dependent variable's judgement performance, fixation duration, and number of fixations. However, no interaction was observed between the two independent variables.

In addition, gender has highly significant impacts on each dependent variable, while context does not significantly affect judgement performance. Considering that this paper aims to recognize motions by the impacts of context, context and fixable duration were selected as the independent variable and dependent variable, respectively, for further discussion.

4.3. Multiple regression model

The SPSS is a nonlinear regression tool to disclose the correlations between significant independent variables and significant dependent variables. In the software, the coefficient of determination R^2 for multivariate correlation reflects the ability of an independent variable to predict the criterion variable, that is, the rate explained by the independent variable. Here, this coefficient is adopted to measure the goodness-of-fit of the regression model formed by the independent variable and the dependent variable, and the explanatory power of the regression equation.

Table 4 presents the details on the cubic regression curve between the independent variable context and the dependent variable fixation duration *(T)*, where $R^2 = 0.780$ and the adjusted $R^2 = 0.796$. The results show that context can explain 79.60% of the variation in the fixation duration. The value of R^2 changed by 0.016 after adjustment.

Table 4. Cubic regression curve between context and fixation duration

R	R Square	Adjusted R^2	Std. Error of the Estimate
0.510	0.780	0.796	680.812

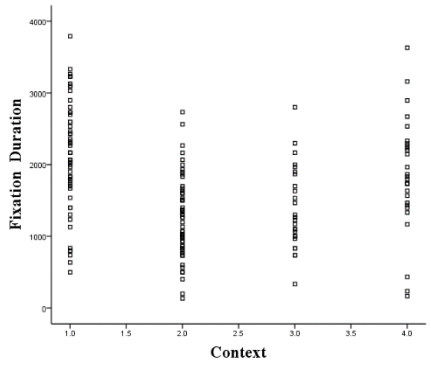

Figure 4. Measured values of different contexts and fixation durations

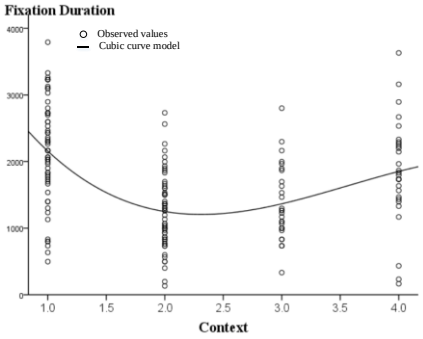

Figure 5. Regression curve between context and fixation duration

The measured values of different contexts and fixation durations are shown in Figure 4, where the 1.0, 2.0, 3.0 and 4.0 on the x-axis respectively stand for "similar motion + target", "similar motion", "mixed motion + target" and "mixed motion". The regression curve between context and fixation duration is displayed in Figure 5. It is clear that the context is nonlinearly correlated with fixation duration. Therefore, the relationship between the two variables was depicted by the cubic regression curve equation below:

$$T = 4743.607 - 3644.881Conte + 116.134Conte^2 - 109.164Conte^3$$

where T is fixation duration; *Conte* is context. This equation measures the actual experimental data and helps to determine the most rational experimental conditions, which lead to the optimal results.

Next, a single factor analysis of variance was performed to further explore the

correlation between context and fixation duration. The regression curve between the two variables was plotted according to the analysis data, and recorded in Table 5 below.

Table 5. Regression curve between context and fixation duration

Context		Mean difference	Std. error	Sig.	95% confidence interval	
					Lower bound	Upper bound
Similar action+ Target (*A*)	Similar action (*B*)	907.625*	128.661	0.000	544.19	1271.06
	Mixed action + Target (*C*)	790.946*	157.577	0.000	345.83	1236.06
	Mixed action (*D*)	304.946	157.577	0.294	-140.17	750.06
Similar action (*B*)	Similar action + Target (*A*)	-907.625*	128.661	0.000	-1271.06	-544.19
	Mixed action + Target (*C*)	-116.679	157.577	0.908	-561.79	328.44
	Mixed action (*D*)	-602.679*	157.577	0.003	-1047.79	-157.56
Mixed action + Target (*C*)	Similar action + Target (*A*)	-790.946*	157.577	0.000	-1236.06	-345.83
	Similar action (*B*)	116.679	157.577	0.908	-328.44	561.79
	Mixed action (*D*)	-486.000	181.955	0.072	-999.97	27.97
Mixed action (*D*)	Similar action + Target (*A*)	-304.946	157.577	0.294	-750.06	140.17
	Similar action (*B*)	602.679*	157.577	0.003	157.56	1047.79
	Mixed action + Target (*C*)	486.000	181.955	0.072	-27.97	999.97
The mean differences are significant at the 0.05 level.						

It can be seen from the curve that the *p* values of factor *A* relative to factors *B* and *C* were both 0.00. Since the *p* values are smaller than 0.05, factor *A* has significant impacts on the latter two factors. Besides, the *p* value of factor *A* relative to factor *D* was 0.294. Since the *p* value is greater than 0.05, factor *A* does not have a significant impact on factor *D*. The *p* value of factor *C* on factor *D* was 0.072. Since the *p* value is smaller than 0.05, factor *C* has a significant impact on factor *D*.

Figure 6 shows the relationship between context and the mean fixation duration. The factors can be ranked as factor *A*, factor *D*, factor *C* and factor *B* in descending order of the fixation duration on different contexts. The comparison between factors *A*, *B* and *C* reveals that the fixation duration in the context increased at the occurrence of the target motion. When the target motion occurred, the participant was able to identify the motion from the context (of similar motions). The comparison between factors *B* and *D* suggests that the fixation duration on the context (of mixed actions) was much longer than that on the pure context (of similar actions) when the target motion did not appear in the situation. The participants were more likely to identify similar motions in the situation in lack of the target motion.

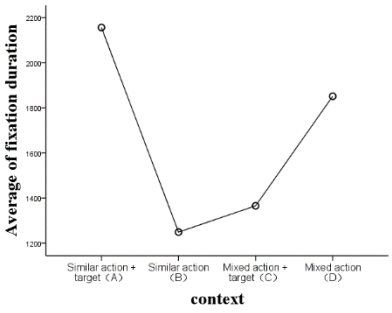

Figure 6. Relationship between context and the mean fixation duration

5. Discussion

Based on the above analysis on eye motion trajectory and fixation duration of the participants, the author put forward a multi-context feature coding strategy to encode virtual situation information.

5.1. Multi-context feature coding strategy

During the coding of visual information, the feature factors in the visual channel were identified to encode the feature information in a manner different from that of the context layer. Firstly, all acquired visual information was registered when the visual feature information was processed in the cognitive system. The expected interval of the feature information was judged. Then, the user motions in, before and

after the interval were taken as similar contexts, and converted into information codes. The multiple sets of visual information were contrasted to remove information with similar codes. After that, the visual information was recoded, sorted and stored in the short-term storage area.

The analysis on multi-context data indicates that the significant correlations only exist among the *A,B* situation, the *A,C* situation, and the *B,D* situation. Hence, the AOIs and eye motion trajectories of the participants were investigated in these three situations. The results are shown in Figure 7 and Figure 8, respectively.

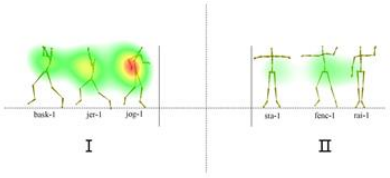

Fig. *a*. *A* * *B* situation

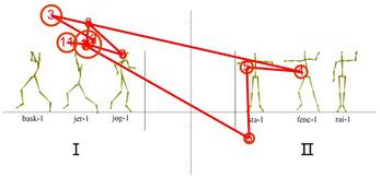

Fig. *d*. *A* * *B* situation

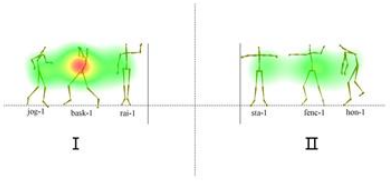

Fig. *b*. *A* * *C* situation

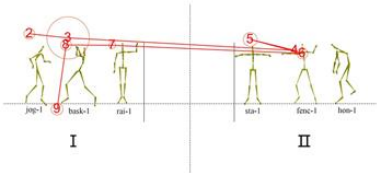

Fig. *e*. *A* * *C* situation

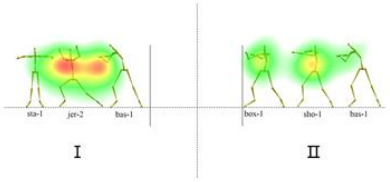

Fig. *c*. *D* * *B* situation

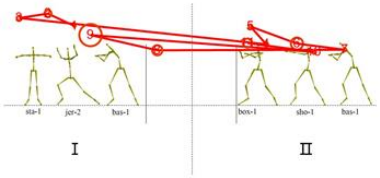

Fig. *f*. *D* * *B* situation

Figure 7. The AOIs in the three situations

Figure 8. Eye motion trajectories in the three situations

Comparing contexts *A* (Figure 7a) and *B* (Figure 8d), it can be seen that the fixation duration was longer on the target motion of context *A* than that of context *B* in the initial phase. This means target motions with similar contexts can be identified easily and highly attractive.

Comparing contexts *A* (Figure 7a) and *C* (Figure 8e), it can be seen that the fixation duration was shorter on the target motion of context *A* than that of context *C*, but the judgement performance in context *A* was better than that in context *C*. The results show that relatively less attention is needed to identify target motion in a similar context.

Comparing contexts *B* (Figure 7c) and *D* (Figure 8f), it can be seen that the fixation

duration and the number of fixations were basically the same in the two contexts. The eyesight of the participant moved back and forth between the two contexts, which means the two contexts have no distinct difference in motion identification. The effect of simple context factors can be eliminated. Since more attention was paid to *D*, the participant must have spent more efforts on recognition of mixed motion than similar motion, that is, the similar motion context can be identified easily.

5.2. Information classification strategy

The feature points of continuous motions were extracted by a Harris 3D detection device and used to characterize user behaviour and identify user motions. The participants' visual information was acquired via computer perception technology, and processed through the perceptual layer, the conceptual layer and the visual layer. This technology was integrated with the information coding mechanism to convert the information into visual codes, which were then sorted in the short term storage area.

Perceptual layer: Based on a multidimensional situation, this layer perceives user information and, together with the user's memory, identifies the acquired information.

Conceptual layer: Based on information classification, this layer processes the registration information at the semantic and conceptual levels. Such information include knowledge and synonymies

Visual layer: Based on information features (e.g. colour, shape and size), this layer classifies and encodes the registration information.

5.3. Information processing strategy

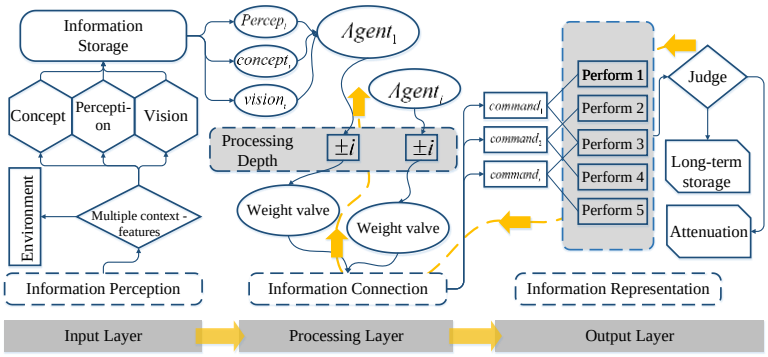

Figure 9. Information processing strategy

The bottom-up processing of visual information is driven by a stimulus occurring at a low cognitive level. The processing results are based on the information observed by the eyes. The feature information is usually represented in the following manner.

After being acquired by the computer, the user information is encoded, classified and stored. The irrelevant information is attenuated, while the relevant information is imported to the perceptual, visual and conceptual channels, creating information images in the input layer. These images are reassembled, and passed up to activate the output layer. Then, the computer executes the instruction information ($command_i$) and saves the feedback information in the long-term storage area. The subsequent similar information will be processed according to the feedback information. As shown in Figure 9, the information processing involves three steps.

Step 1: The current user information is perceived and converted quickly into codes. The codes are entered into the short-term storage area, and reclassified and encoded according to the difference in information features. The mapping between the information in different channels and the computer resources is established to complete the information encoding on the input layer.

Step 2: The information image of the processing layer is activated by the information from the input layer. The information of the activated image is reclassified and passed up as information instructions. Comparing each eigenvector with multiple other eigenvectors, each pair of information images is connected based on the mapping between them. The information is concatenated once, a signal of the storage of an execution event. After the activation of an information instruction, the PD and connection weight increase by a certain degree. Once the weight value reaches the threshold, the instruction information is triggered and passed up to the output layer.

Step 3: On entering the output layer, the information instruction is executed to represent the information. The user will complete the operation according to the represented information. If the represented information satisfies the user's operation demand, the information will enter the long-term storage, marking the end of information processing. Meanwhile, the information will also be passed back to the input layer, and the corresponding PD will be increased by one unit. Otherwise, the PD will be reduced by one unit.

6. Conclusions

This paper designs a cognitive experiment for user behaviours in virtual scene and explores the influence of different contexts on behaviour recognition. Through the analysis of the experimental data, a nonlinear relationship was discovered between the fixation duration and context. In addition, interaction occurred between different contexts during the recognition process. The recognition efficiency was higher at the appearance of the target motion with the context (of similar motions) than at that of the target motion with the mixed context. Based on the experimental data, the author designed strategies for visual information feature coding, information classification and information processing.

Of course, there are several limitations in this research. For instance, actual human motion was not covered in the cognitive experiment due to limited experimental conditions. The experimental results should be closer to actual data in the virtual scene, if videos of complex background and human motion are examined in the experiment.

Besides, the theoretical coding model should be further tested on actual virtual interactive platforms and other cases. As a result, future research will detail the information retrieval and processing methods in virtual situations, and study the behavioural and cultural connotations of virtual situations using motion capture and eye motion tracking system.

Acknowledgments

Qingsheng Xie is the corresponding author.The work is supported by Guizhou Province Science and Technology Fund Project (Guizhou Branch Support [2017] 2870;[2016]2327).Guizhou Province Education Department Talent Project(Guizhou Branch Education KY [2017]062). GuizhouHuman Resources and Social Security Bureau (Guizhou Human Project Fund [2016] 06). Guizhou Province Science and Technology Fund Project (Guizhou Branch Total LH[2017]1046).

Reference

Almaev T. R., Valstar M. F. (2013). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356–361. https://doi.org/10.1109/ACII.2013.65

Althloothi S., Mahoor M. H., Zhang X., Voyles R. M. (2014). Human activity recognition using multi-features and multiple kernel learning. *PATTERN RECOGN*, Vol. 47, pp. 1800-1812. https://doi.org/10.1016/j.patcog.2013.11.032

Barbosa I. B., Cristani M., Bue A. D. (2012). Re-identification with RGB-D sensors. *International Conference on Computer Vision. Springer-Verlag*, pp. 433-442. https://doi.org/10.1007/978-3-642-33863-2_43

Bilen H., Fernando B., Gavves E., Vedaldi A., Gould S. (2016). Dynamic image networks for action recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3034–3042, https://doi.org/10.1109/CVPR.2016.331

Bobick A. F., Davis J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell*. Vol. 23, pp. 257-267. https://doi.org/10.1109/34.910878

Chen B., Chen W. (2011). Noisy image segmentation based on wavelet transform and active contour model. *APPL ANAL*, Vol. 90, pp. 1243-1255. https://doi.org/10.1080/00036811003717939

Chen C., Li Y. Q., Liu W., Huang J. Z. (2016). Image fusion with local spectral consistency and dynamic gradient sparsity. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2760-2765. https://doi.org/10.1109/CVPR.2014.347

Cheng S., Hsu C., Li J. (2013). Combined hand gesture — speech model for human action recognition. *SENSORS-BASEL*, Vol. 13, pp. 17098-17129. https://doi.org/10.3390/s131217098

Fr D., Landragin R. (2006). Visual perception, language and gesture: a model for their understanding in multimodal dialogue systems. *SIGNAL PROCESS*, Vol. 86, pp. 3578-

3595. https://doi.org/10.l016/j.sigpro.2006.02.046

Hemati R., Mirzakuchaki S. (2014). Using local-based harris-PHOG features in a combination framework for human action recognition. *Arab J Sci Eng*, Vol. 39, No. 2, pp. 903-912. https://doi.org/10.1007/s13369-013-0816-6

Ji S. W., Xu W., Yang M., Yu K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 35, No. 1, pp. 221–231. https://doi.org/10.1109/tpami.2012.59

Jiang Z., Lin Z., Davis L. S. (2012). Recognizing human actions by learning and matching shape-motion prototype trees. *Ieee T Pattern Anal*, Vol. 34, pp. 533-547. https://doi.org/10.1109/tpami.2011.147

Lehrmann A. M., Gehler P. V., Nowozin S. (2014). Efficient nonlinear markov models for human motion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1314–1321. https://doi.org/10.1109/CVPR.2014.171

Loncomilla P., del Solar J. R., Martnez L. (2016). Object recognition using local invariant features for robotic applications: A survey. *Pattern Recognit*., Vol. 60, pp.499-514. https://doi.org/10.1016/j.patcog.2016.05.021

Maji S., Bourdev L., Malik J. (2011). Action recognition from a distributed representation of pose and appearance. *In, Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition: IEEE Computer Society*, pp. 3177-3184. https://doi.org/10.1109/CVPR.2011.5995631

Matikainen P., Hebert M., Sukthankar R. (2010). Representing pairwise spatial and temporal relations for action recognition. *In: Daniilidis K, Maragos P, Paragios N eds, Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I. Berlin, Heidelberg: Springer Berlin Heidelberg*, pp. 508-521. https://doi.org/10.1007/978-3-642-15549-9_37

Scovanner P., Ali S., Shah M. (2007). A 3-dimensional sift descriptor and its application to action recognition. *Proceedings of the Fifteenth International Conference on Multimedia*, pp. 357–360. https://doi.org/10.1145/1291233.1291311

Shotton J., Fitzgibbon A., Cook M. (2011). Real-time human pose recognition in parts from single depth images. *Computer Vision and Pattern Recognition. IEEE*, pp. 1297-1304. https://doi.org/10.1145/2398356.2398381

Slama R., Wannous H., Daoudi M., Srivastava A. (2015). Accurate 3D action recognition using learning on the grassmann manifold. *Pattern Recognit*. Vol. 48, No. 2, pp. 556–567. https://doi.org/10.1016/j.patcog.2014.08.011

Valstar M. (2015). Automatic facial expression analysis, understanding facial expressions in communication. *Springer India*, pp. 143-172. https://doi.org/10.1007/978-81-322-1934-7_8

Wang P., Li W., Ogunbona P., Gao Z., Zhang H. (2014). Mining mid-level features for action recognition based on effective skeleton representation. *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*, pp. 1–8. https://doi.org/10.1109/DICTA.2014.7008115

Zanfir M., Leordeanu M., Sminchisescu C. (2013). The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. *Proceedings of the*

IEEE International Conference on Computer Vision, pp. 2752–2759. https://doi.org/10.1109/ICCV.2013.342

Zhang H., Zhou W., Parker L. E. (2014). Fuzzy segmentation and recognition of continuous human activities. *IEEE International Conference on Robotics and Automation*, pp. 6305-6312 https://doi.org/10.1109/ICRA.2014.6907789

Zhang J., Lin H., Nie W. (2015). Human action recognition bases on local action attributes. *J Electr Eng Technol*, Vol. 10.

Zhu Y. D., Kikuo F. (2010). A bayesian framework for human body pose tracking from depth image sequences. *Sensors-Basel*, Vol. 10, pp. 5280-5293. https://doi.org/10.3390/s100505280