Presentation Training System Based On Imitating Past Famous Speech

Auther : SHI Yuhua

Student Number : 6612160065-0

Supervisor : Prof. Haruo Noma

Date : 2018. 7. 18

Ritsumeikan University

Graduate School of Information Science and Engineering

Abstract

ページ以内にしましょう．Intro と Concl からの文章のコピーは NG.（同じ文があると手抜きと思われます.）

CONTENTS

List of Figures

List of Tables

1. Introduction

The presentation is the art of persuasion, It plays a significant role in our society and has the tremendous impact on the success of everyone [1]. The presentation commonly used to communicate the presenter's ideas to the listener. However, giving a presentation successfully like John F. Kennedy or Steven Jobs is not a simple thing. A great preparation not only contains verbal style but also needs various nonverbal behaviors. On one hand, the content of a preparation must be clear, vivid and appropriate[2]. On the other hand, the significant component of a presentation lies upon nonverbal cues which have the power to change the meaning assigned to spoken words[1].

Nonverbal behaviors of public speakers are expressed via several channels such as voice, gesture and facial expression. They have been proven to have a more significant influence than verbal cues. According to Argyle nonverbal messages are thirteen to fourteen times more powerful than verbal ones[3]. Likewise, Arcy showed that the audience receives more than half of information from nonverbal behaviors[4]. The study of Seiler[1]shows the fact that most people unconsciously believe more in nonverbal behaviors than verbal cues.

Unfortunately, it's hard to practice to express the effective nonverbal behaviors because they are mostly expressed subconsciously. To achieve the learning results, trainees must be provided with appropriate feedbacks from human advisors, but it is costly and not always available. According to Seiler [1], imitation is a kind of effective method to help the trainee to refine their nonverbal behaviors. Imitating those past famous speakers' gesture, sound or enunciation to learn what they're doing right, and then try to make it your own as Pablo Picasso said *" Good Artists Copy. Great Artists Steal "*. It is a great learning experience and can stretch trainees' abilities.

In parallel, the role of nonverbal behaviors in computing is becoming increasingly recognized by the development of the emerging fields, such as social signal processing and affective computing. Such as [5–9], those Deep Learning library can extract user's body key point with high accuracy. Therefore, computers have been equipped with the abilities to decode the complexity of humans nonverbal channels.

Many papers discussed some approaches toward the automatic recognition of nonverbal from trainees. However these approaches all defined some exact rules in advance to give a score of

the presentation, but few have focused on imitating past famous speech to refine their nonverbal behaviors.

In this paper, we propose a presentation training system that allows the trainee to imitate past famous speech to improve their nonverbal behaviors. Nonverbal behaviors include many aspects, and we choose to analyze the gesture of orators as the nonverbal behavior in this paper. In advance, we employed OpenPose library[6] to extract orators' motion data from past famous speech 2D video. While training, the system capture the trainee's motion in real-time using Microsoft Kinect. We choose the cosine similarity of adjacent limbs as the feature to calculate the score that shows the similarity of the trainees' motion and the motion of extracted orators. The trainees will also wear an HMD that show a virtual hall and some virtual audience. The audience will perform some actions according to the score.

Figure 1.1 : System Introduction

The rest of this paper is organized as follows. First we will introduce some existing presentation training method and proposed solution in the chapter 2. Then we will introduce the preparation about OpenPose, Kinect and the Evaluation of presentation in the chapter3. The chapter 4 describes our proposed system in detail. In the chapter 5, we will evaluate the system and discuss the results. The last chapter is for conclusions and future works.

2. Presentation Training Method and Related Work

2.1 Presentation Training Method

Presentation skills give us the power to change the world. Great presenters instill trust, engage our minds and hearts, deliver ideas and information, and inspire and captivate us.

2.1.1 balabala

2.2 Related Work

In this paper, we propose a presentation training system that allow trainees to imitate past famous speech to improve their nonverbal behaviors. Many papers discussed some approaches toward the automatic recognition of nonverbal from trainees. However these approaches all defined some exact rules in advance to give a score of the presentation, but few have focused on imitating past famous speech to refine their nonverbal behaviors. Some researches analyze some visual and vocal channels of trainees, thus can provide information about their presentations. For example, the system in Pfister was originated from a vocal emotion detection module [10]. It was similar to the approach of Silverstein [11], which was built solely on vocal cues, by analyzing the voice such as pitch or tempo. The Hincks's system measured the changes in vocal pitch, and then give visual feedbacks to promote pitch variation by relying on the importance of pitch variance in oral presentations [12].

On the other hand, some include nonverbal behaviors in the analysis. Gao introduced the method based only on visual information [13]. In contrast, Kurihara added face position and orientation as the approximation of eye contact, together with pitch, speaking rate, utterance and filled pauses [14]. Hoque proposed a automated conversation coach to help trainees improve their interview skills by recognize their motion and facial expression [14].

We will introduce the detail about Kurihara's research [14] in section 2.2.1 and Nguyen's research [15] in section 2.2.2.

2.2.1 Presentation Sensei

In Kurihara's paper they present a presentation training system that observes a presentation rehearsal and provides recommendations for improving the delivery of the presentation, such as to speak more slowly and to keep eye contact with the audience (Figure 2.1). Kurihara intentionally focus on the basic behavior patterns because high-level semantics is very difficult to analyze by computers in contrast to basic behavior patterns. Kurihara's goal is to help people improve their base-line presentation skills by reducing inappropriate behavior such as using many fillers (e.g. " er ") or continuously looking down at the script.

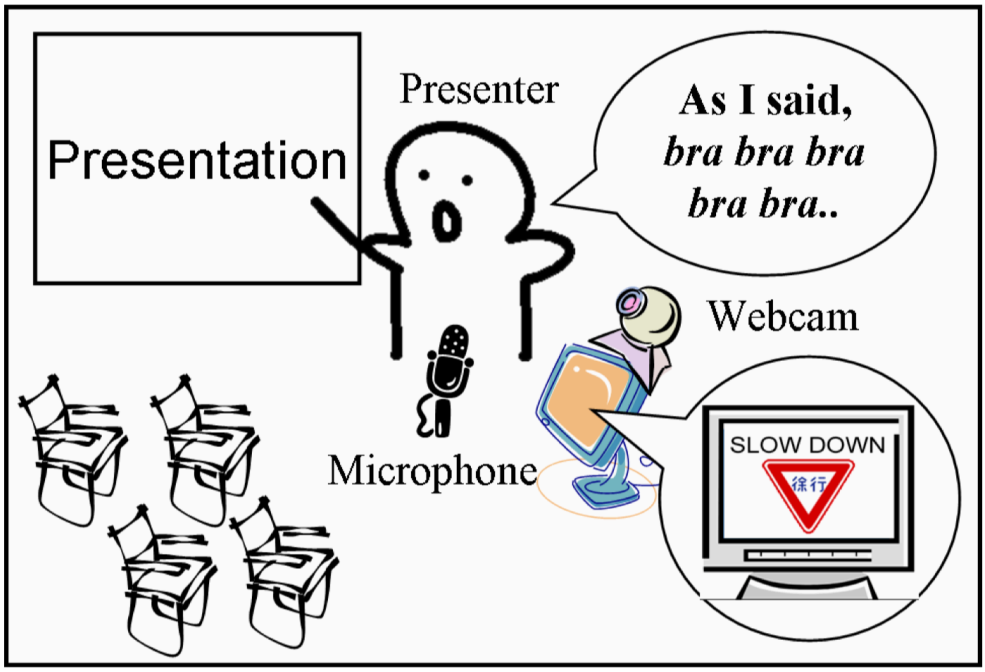

Figure 2.1 : Presentation sensei system

In this work, Kurihara focus on the following five aspects of presentation delivery.

- The speaking rate should not be too fast.
- The speech should not be monotonous.
- The speech should not contain too many fillers.
- The speaker should look at the audience and avoid continuously looking down at a script or a screen.
- The speaker should finish the presentation within a certain time limit.

Kurihara selected these because they are emphasized in existing literature and they can be detected to some extent using current speech processing and image processing technologies. The Presentation Sensei system visualizes the analysis result in real time communicating with a presentation tool. It can give the presenter both instant “ online ” feedback and post “ offline ” feedback for improvements. The online feedback function shows the analysis result in real-time. When the system detects some inappropriate behavior, it alerts the presenter by showing a visual signal (Figure 2.2). The offline feedback function shows the visual summaries of the indices for the presenter ’ s self-examinations (Figure 2.3).

Figure 2.2 : Online feedback. (Left) Real time monitor. (Trafic signals) Visual Alerts.

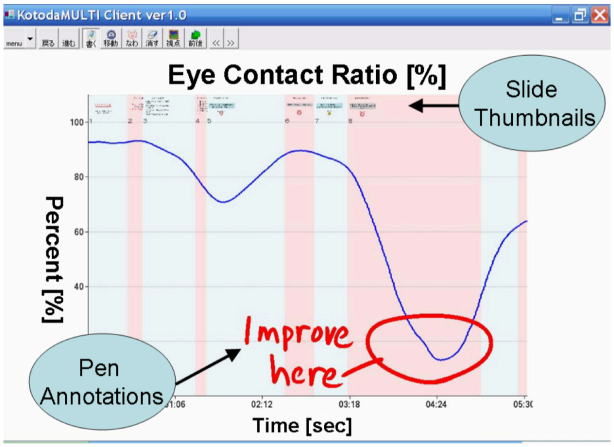

Figure 2.3: Offline feedback. An example of the generated charts The presenter can annotate them with a pen.

It is often difficult to make speech processing and image processing work robustly in adverse environments. One advantage of Kurihara’s target application domain, personal presentation rehearsal, is that we can relatively freely configure the environment. It is realistic to rehearse in a silent room with no visual obstacles, where these technologies perform the best. This feature makes the Presentation Sensei system a highly practical application even with imperfect recognition technologies [14].

Kurihara's system is unique in that, while general multimodal systems help the user to control computers, it tries to help computers guide humans. This way of using a computer is relatively new. Heer et al. [16] investigated the design guidelines for this sort of systems and also introduced an experimental media capture system that acts as a film director. The contribution of Kurihara's work is to introduce a practical application based on this approach and to show its feasibility using state-of-the-art speech and video recognition technologies [14].

System Configuration

The Presentation Sensei system consists of several modules connected by a network (Figure 2.4). The audio analysis module continuously analyses the input signal from a microphone and provides the integration module with the results of the utterance duration detection, pitch (F0) detection, and filled pause detection. The speech recognition module also continuously provides the integration module with the mora-based speech recognition results. The image processing module continuously analyzes the input from a webcam and provides the integration module with the result of the face position/orientation detection. The integration module integrates all the provided information and gives feedback to the presenter using various monitors. These modules can be distributed over a local network for the load sharing purpose and they communicate with each other via RVCP protocol [17]. The system can also be connected to a third party presentation tool to achieve synchronization. Kurihara currently connect their system to an in-house presentation program to receive timing information and thumbnail images of slides.

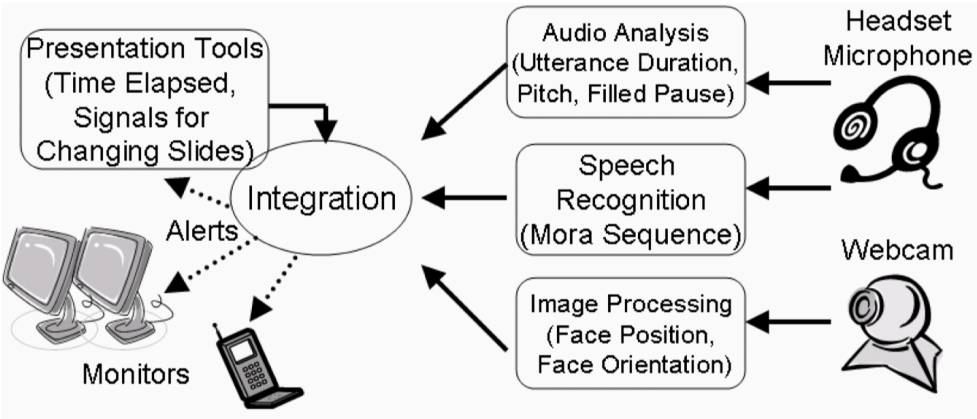

Figure 2.4 : System configuration of presentation sensei

3. Preparation

In this chapter, we will introduce some preparation work for the proposed system. At first, we employed the OpenPose library [6] to extract the orator's joint data from past speech 2D video, and we will introduce the OpenPose library in section 1. Then we set up a Microsoft Kinect for Windows Version device [18] to extract the trainee's joint data in training, and we will introduce the Kinect camera in section 2. To evaluate the effectiveness of the proposed system, we need to know how to evaluate a presentation, and we will introduce some evaluation points of a presentation.

3.1 Joint Data Extraction Method (2D)

We employed the OpenPose library to extract the orator's joint data from 2D speech video [6]. OpenPose is a library for real-time multi-person keypoint (Figure 3.1) detection and multi-threading written in C++ using OpenCV and Caffe [19]. OpenPose can detect human body, hand and facial keypoionts on single images. In addition, the system computational performance on body keypoint estimation is invariant to the number of detected people in the image.

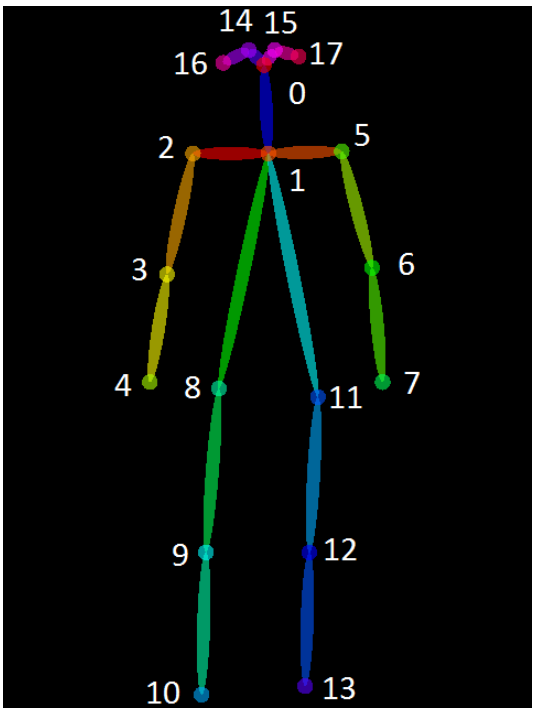

Figure 3.1 : Keypoints detected by OpenPose

Convolutional Pose Machine (CPM)

Convolutional Pose Machine (CPM) use Convolutional Neural Networks (CNNs) to detect the human joint data from 2D image or video. CPMs consist of a sequence of convolutional networks that repeatedly produce 2d belief maps for the location of each part (Figure 3.2). At each stage, image features and belief maps produced by the previous stage are used as input [20].

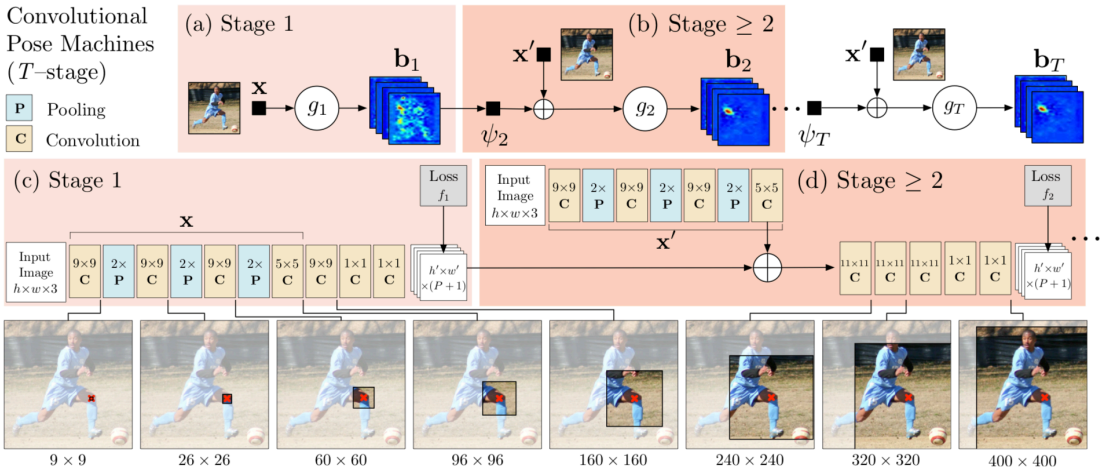

Figure 3.2 : Architecture of Convolutional Pose Machines (CPMs)

The belief maps provide the subsequent stage an expressive non-parametric encoding of the spatial uncertainty of location for each part, allowing the CPM to learn rich image-dependent spatial models of the relationships between parts. The overall proposed multi-stage architecture is fully differentiable and therefore can be trained in an end-to-end fashion using back propagation [20].

At a particular stage in the CPM, the spatial context of part beliefs provides strong disambiguating cues to a subsequent stage. As a result, each stage of a CPM produces belief maps with increasingly refined estimates for the locations of each part.

In order to capture long-range interactions between parts, the design of the network in each stage of our sequential prediction framework is motivated by the goal of achieving a large receptive field on both the image and the belief maps [20].

OpenPose

Based on CPM architecture, OpenPose is an efficient method for multi-person pose estimation what uses a non-parametric representation of association scores via Part Affinity Fields (PAFs), a

set of 2d vectors field that encodes the location and orientation of limbs over the image domain.

The part affinity is a 2d vector field for each limb for each pixel in the area belonging to a particular limb, which encodes the direction that points from one part of the limb to the other. Each type of limb has a corresponding affinity field jointing its two associated body parts. A greedy parsing algorithm is sufficient to produce high-quality parses of body poses, which maintains efficiency even as the number of people in the image increase [6].

Figure 3.3 illustrates the overall pipeline of OpenPose library.

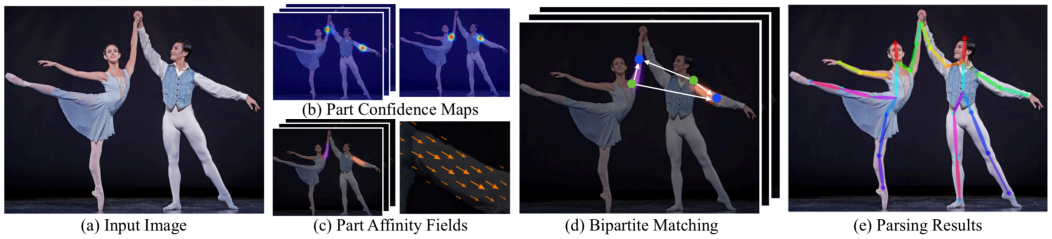

Figure 3.3 : Overall pipeline of OpenPose

- OpenPose takes, as input, a color image of size of $w \times h$ (Fig 3.3a) and produces, as output, the 2d locations of anatomical key-points for each person in the image (Fig 3.3e) [6].
- First, a feed-forward network simultaneously predicts a set of 2d confidence maps of body part locations (Fig 3.3b) and a set of 2d vector fields of part affinities, which encode the degree of association between parts (Fig 3.3c).
- Finally, the confidence maps and the affinity fields are parsed by greedy inference (Fig 3.3d) to output the 2d keypoints for all people in the image.

The architecture of OpenPose, shown in Figure 3.4, simultaneously predicts detection confidence maps and affinity fields that encode part-to-part association. The network is split into two branches: the top branch, shown in beige, predicts the confidence maps, and the bottom branch, shown in blue, predicts the affinity fields. Each branch is an iterative prediction architecture, Following the typical structure of a CMP, which refines the predictions over successive stages, with intermediate supervision at each stage.

In our proposed system, we employed OpenPose library to extract the orator's joint data from past famous 2D speech video (Fig 3.5). We use those joint data to do template matching to get the score.

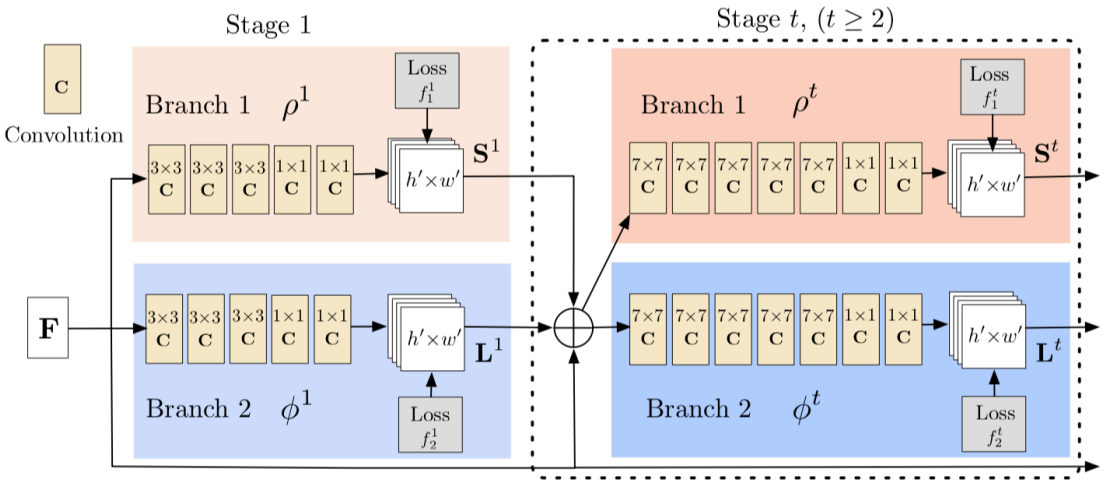

Figure 3.4：Architecture of the two-branch multi-stage CNN in OpenPose

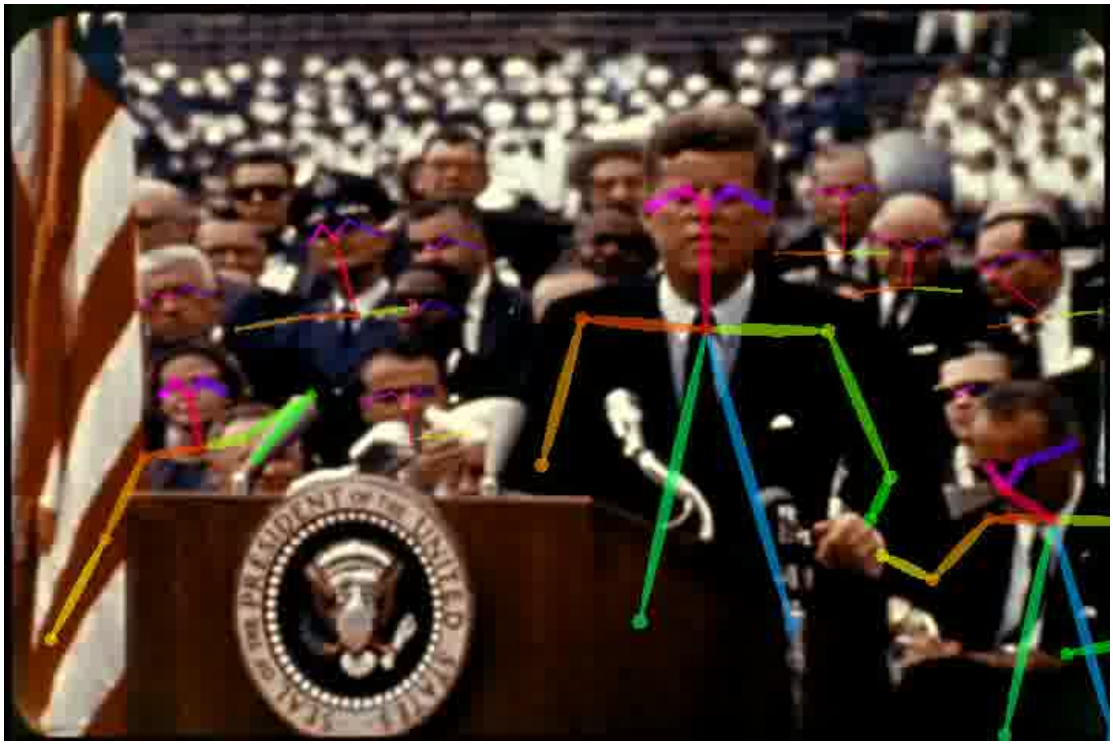

Figure 3.5：Joint detected by OpenPose

3.2 Joint Data Extraction Method (3D)

Recent advances in 3D depth cameras such as Microsoft Kinect sensors have created many opportunities for multimedia computing. The Kinect sensor lets the computer directly sense the third dimension (depth) of the players and the environment. It also understands when users talk, knows who they are when they walk up to it and can interpret their movements and translate them into a format that developers can use to build new experiences [21].

Kinect Sensor

The Kinect sensor incorporates several advanced sensing hardware. Most notably, it contains a depth sensor, a color camera, and a four-microphone array that provide full-body 3D motion capture, facial recognition, and voice recognition capabilities (see Figure 3.6).

Figure 3.6b shows the arrangement of the infrared (IR) projector, the color camera, and the IR camera. The depth sensor consists of the IR projector combined with the IR camera, which is a monochrome complementary metal oxide semiconductor (CMOS) sensor. Although the exact technology is not disclosed, it is based on the structured light principle. The IR projector is an IR laser that passes through a diffraction grating and turns into a set of IR dots [21].

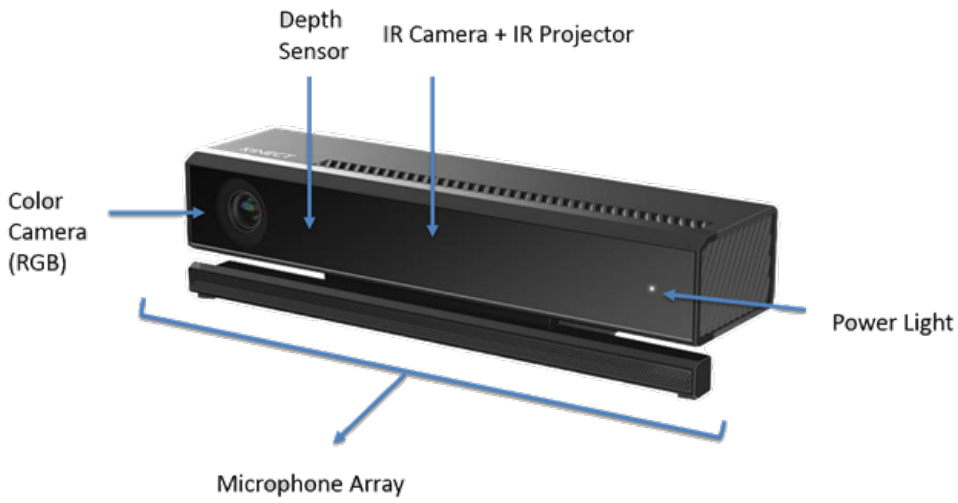

Figure 3.6：Microsoft Kinect Sensor

The relative geometry between the IR projector and the IR camera as well as the projected IR dot pattern are known. If we can match a dot observed in an image with a dot in the projector pattern, we can reconstruct it in 3D using triangulation. Because the dot pattern is relatively ran-

dom, the matching between the IR image and the projector pattern can be done straightforwardly by comparing small neighborhoods using, for example, normalized cross correlation [21].

Kinect Skeletal Tracking

In skeletal tracking, a human body is represented by some joints representing body parts such as head, neck, shoulders, and arms (see Figure 3.7). Each joint is represented by its 3D coordinates.

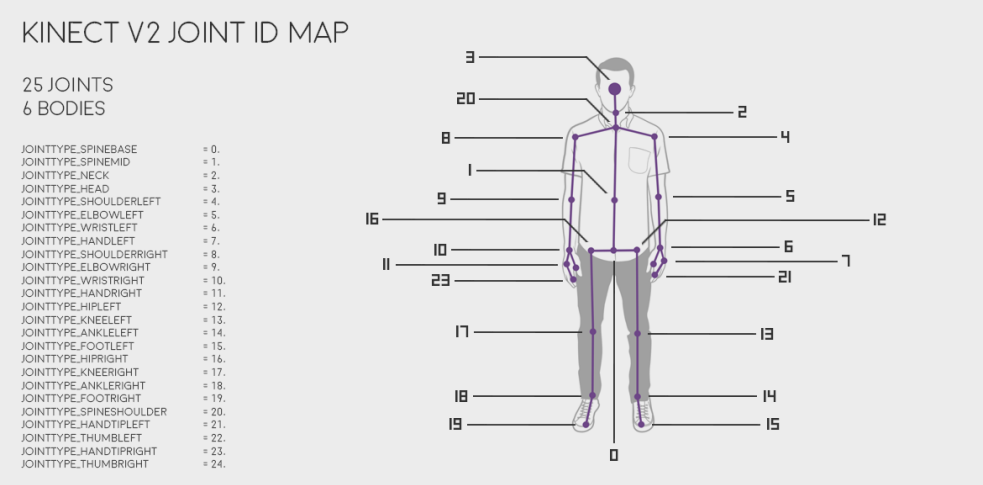

Figure 3.7：Microsoft Kinect V2 Joint Id Map

Figure 3.8 illustrates the whole pipeline of Kinect skeletal tracking. The first step is to perform per-pixel, body-part classification. The second step is to hypothesize the body joints by finding a global centroid of probability mass through the mean shift. The final stage is to map hypothesized joints to the skeletal joints and fit a skeleton by considering both temporal continuity and prior knowledge from skeletal train data.

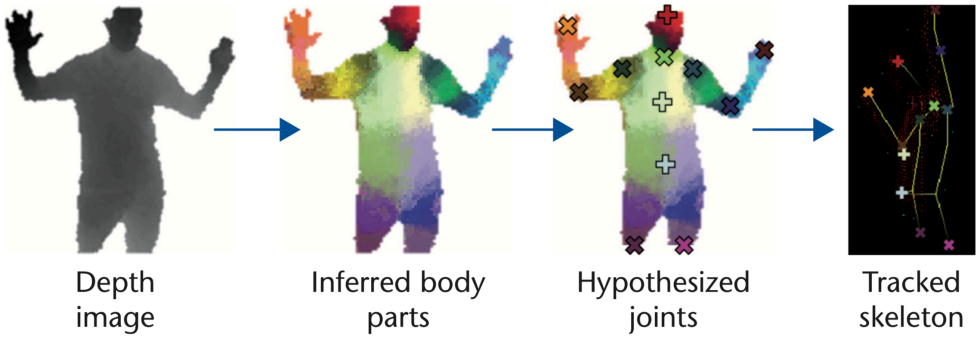

Figure 3.8：The Kinect skeletal tracking pipeline

In our proposed system, we set up a Kinect V2 to extract the joint data of trainee in real-time. We only employ the 2D coordinates to fit the joint data extracted by OpenPose from the past speech.

3.3 Evaluation of presentation

To evaluate the presentation skill of trainees, we need to know what kind of behaviors will have the impact on the presentation. Nguyen's research performs an observation to analysis the behaviors of presenters. They collected data from a training class about public speaking skills for postgraduate students. They ask the learners to give short presentations (about one minute) in front of the audience, which includes about ten other learners and one or two coaches. The presenters can freely choose the content of the presentations. In fact, all presenters chose to talk about their research, in the ways that it can be understood by all of the audience that might come from the different fields. After each presentation, the audience gave feedbacks and suggestions on how the presentation should be improved, regarding nonverbal expressions. They set up a regular camera to record the presentations. They also set up a Microsoft Kinect to capture the whole body movement for their further signal processing, as well as behavioral studies. They stored the data from Kinect as the *.ONI files using the OpenNI SDK. They removed the unsatisfied videos (e.g. presenters moved out of the camera range) and finally collected 39 presentations of 11 presenters (four females and seven males) [15].

In their research, they use regular videos for behavioral analysis. This task was done through the collaboration with an expert in public speaking. The role of the expert was to review the recorded videos, and then specifying the nonverbal cues that affected the performance of the speakers, together with the durations that they appeared. Thus, for each video, a set of behaviors was created. They collected the nonverbal cues and then annotated their appearance using the commercial software Noldus Observer XT [22]. Behaviors were categorized into either *State event* if their duration is necessary to be studied, or *Point event* otherwise. The software provided them with the statistical analysis on the appearance of these behaviors, including the number of presentations that contain the behaviors, the rate that they appeared (point events) and the percentage of time that they accounted for (Table 3.1) [15].

Table 3.1 : **The list of observed nonverbal cues**

#	Behaviors	Event Type (S/P)	No.	Rate of occurrences (times/minute)			Percentage during observation of the occurrences(%)		
				M	SD	Range	M	SD	Range
Postural behaviors									
1	(-) Shoulder too tight	S	19				60.94	23.80	12.67 - 98.50
2	(-) Legs closed	S	12				73.02	36.44	5.15 - 100
3	(-) Legs too stretch	S	3				61.42	11.33	19.18 - 100
4	(-) Weight in on foot	S	20				65.42	28.69	5.20 - 100
5	(-) Chin too high	S	14				64.94	23.80	12.67 - 98.50
6	(-) Hands in pockets	S	3				11.85	4.89	12.76 - 92.60
7	(+) Lean forward	S	19				32.50	28.66	3.70 - 82.78
8	(-) Lean backward	S	17				62.80	28.07	12.73 - 96.20
Vocal behaviors									
9	(-) Speak too fast	S	19				45.88	36.55	7.32 - 100
10	(-) Start too fast	P	18						
11	(-) Energy decreases at the end	P	23	2.88	1.77	0.53 - 6.31			
12	(+) Vocal emphasis	P	33	5.51	4.51	0.59 - 17.50			
13	(+) Suitable pause	P	33	4.63	3.16	0.53 - 12.50			
14	(-) Unsuitable pause	P	20	1.73	1.14	0.53 - 5.19			
15	(-) Monotone	S	20				92.49	13.08	56.29 - 100
16	(-) Fillers	P	34	5.17	4.22	1.44 - 19.03			
17	(-) Stuttering	P	12	1.72	0.83	0.53 - 3.42			
Behaviors of eye contact									
18	(-) Make eye contact	S	39				93.81	8.24	75.00 - 100
19	(-) Contact avoidance	S	28				9.98	8.47	1.12 - 25.00
19.1	(-) Look up to ceiling	S	14				4.23	2.95	1.12 - 9.61
19.2	(-) Look down to floor	S	19				7.67	4.67	2.84 - 14.17
19.3	(-) Look at hands	S	11				10.24	3.15	4.40 - 13.15
Behaviors related to facial expression									
20	(+) Facial mimicry	S	30				39.31	25.97	4.50 - 91.81
21	(-) Smile	S	22				13.62	11.54	3.54 - 41.08
22	(-) Flat face	S	8				80.61	24.16	40.41 - 100
Behaviors related to whole body movement									
23	(-) Too much movement	S	11				42.21	25.97	4.50 - 91.81
24	(-) Too little movement	S	23				50.62	29.21	10.05 - 100
25	(-) Step backward	P	31	1.83	1.27	0.36 - 4.36			
26	(+) Step forward	P	34	2.06	1.04	0.59 - 4.61			
Behaviors related to hand gesture									
Amount of hand gesture									
27	Hand gesture occur	P	38	16.83	7.15	0.93 - 28.42			
28	(-) Too little gestures	S	20				69.55	34.64	17.21 - 100
29	(-) Too much gestures	S	10				61.49	31.82	27.34 - 96.10
Quality of hand gestures									
30	(-) Bounded gestures	P	30	6.75	5.33	1.00 - 19.77			
31	(+) Relaxed gestures	P	29	7.41	4.95	1.15 - 15.79			
32	(-) Casual gestures	P	10	5.16	3.14	1.56 - 10.28			
33	(-) Uncompleted gesture	P	27	3.23	2.78	0.93 - 10.27			
34	(+) Gestural emphasis	P	20	4.43	4.05	0.36 - 11.99			
35	(-) Repeated gestures	P	31	6.57	2.49	1.09 - 12.31			

The observed behaviors can be separated based on the nonverbal channels that they were generated: (1) Posture (the static configuration of body), (2) Voice (concerning the paralinguistic characteristics), (3) Eye contact, (4) Facial Expression, (5) Globe body movement, (6) Hand gesture. This method of categorization is similar to the literature of public speaking skills [2]. From their observation, as well as advices from the expert, the following aspects are the most important:

- *Eye Contact* : Similar to social interaction, maintaining good eye contact is the first thing the presenters must keep in mind. It initiates and strengthens the connection between them and the audience (#18, 19 in Table 3.1). It might have the first and foremost influence to the performance of a presentation, as well as regular communications [22].

- *Amount of energy* : This aspect concerns the dynamic characteristics of a presentation, thus can reflect the internal state of the presenters. It has impact in most behaviors that they have found (except posture as the static channel). For example, the amount of whole body movement (#23, 24), the amount of hand gesture (#28, 29), vocal behaviors (partly via tempo, emphases) and most features of hand gesture [15].

- *Variety* : The presentations with strong variations significantly increase the attention of the audience. Lacking variation results in monotone (#15), flat face (#22), and hand gesture repeated (#35). In fact, variety can be separated as one single measurement to analyze a presentation. It takes the role as rhythm in music. Even a beautiful piece of music, without changes in rhythm will steadily lose attention from the audience [15].

In order to evaluate the effectiveness of our proposed system, we selected 21 import cues from Nguyen's research [15] and did an experiment. We make a evaluate sheet like table 3.2 to evaluate the presentation skills of the trainees before and after the training. We will explain the details about our experiment in chapter 5.

Table 3.2：Nonverbal cues for evaluation

#	Behaviors	Event Type (S/P)
Postural behaviors		
1	(-) Hands in pockets	S
2	(+) Lean forward	S
3	(-) Lean backward	S
Vocal behaviors		
4	(-) Speak too fast	S
5	(+) Vocal emphasis	P
6	(+) Suitable pause	P
7	(-) Unsuitable pause	P
Behaviors of eye contact		
8	(-) Make eye contact	S
9	(-) Contact avoidance	S
10	(-) Look up to ceiling	S
11	(-) Look down to floor	S
Behaviors related to facial expression		
12	(-) Smile	S
13	(-) Flat face	S
Behaviors related to whole body movement		
14	(-) Too much movement	S
15	(-) Too little movement	S
16	(-) Step backward	P
17	(+) Step forward	P
Behaviors related to hand gesture		
18	(+) Hand gesture occur	P
19	(-) Too little gestures	S
20	(-) Too much gestures	S

4. Proposed Presentation Training System

4.1 Overview of System

4.2 Extract Pose Data from Past Speech Video

4.3 Extract Pose Data from Trainees

4.4 Template Matching Method

4.5 Feedback for speech

5. Evaluation of the System Effectiveness

本章では，提案した〇〇手法を〇〇に適用した実験結果とその考察を示す.

5.1 Method

5.2 Results

5.3 Discussion

Table 5.1 : My caption

#		1st			2nd		
	Evaluator number	1	2	3	1	2	3
Postural behaviors							
-1	Hands in pockets	1					
-1	Lean backward						
1	Lean forward	1					1
Whole body movement							
-1	Too much movement						
-1	Too little movement	1	1	1			
-1	Step backward						
1	Step forward						
Vocal behaviors							
-1	Speak too fast						
-1	Unsuitable pause			7			2
1	Suitable pause	1			2	1	
1	Vocal emphasis	1					
Behaviors of eye contact							
1	Make eye contact				3	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling		1				
-1	Look down to floor						1
Facial expression							
1	Smile				3		1
-1	Flat face	1	1	1			
Hand gesture							
1	Hand gesture occur		2		5	5	1
-1	Too little gestures	1		1			
-1	Too much gestures						
Point of each evaluator		-2	-2	3	13	7	1
Average Point		-5			7		

6. Conclusion

本研究では，○○という手法において，○○保存する命令を，分岐の信頼性を用いて限定する手法を提案した.（略）その結果，再利用できない無駄な命令の保存を，最大で95%削減することができた．今後は，さらに○○の部分を改良することにより削減割合をさらに向上させることが課題である.

ここまできたら，力尽きそうですが，力を振り絞って，

- 本論文の概要と特徴
- 得られた成果
- それから得られる最終結論
- 残された課題

は書いてください.

Acknowledgement

本論文の執筆にあたり，多岐に渡るご指導をしてくださり，私を導いてくださった野間春生先生、Lopez 先生、松村耕平先生に深く感謝の言葉を申し上げます. 至らぬ私が無事に論文の執 筆を終えるができたのは Lopez 先生と野間春生先生の教えがあったからです. 本当にあ りがとうございます. また，副指導として鋭い指摘や温かい助言を下さった周勇先生に深 く感謝の意を述べさせて頂 x ます. 中間審査において私が見落としている点をご指摘いた だき，今後の方向性についてご提案いただきました李亮先生に心から感謝致します. 研究 の日々を互いに切磋琢磨し，励まし合いながら過ごした研究室の皆様に輝かしい未来があ るようお祈り申し上げます. 自らもお忙しいなか，快く実験への協力を承諾いただきまし た被験者の皆様への恩は一生涯胸に残します. 研究のみならず数々の困難を共に乗り越え てきた同級生の方々が，これからも健やかに過ごし，皆様の力を存分に発揮できるような 日々を送られますよう心からお祈りいたします. 色々なリスクを顧みずに大学院進学への 許可を快く下さった両親に心から感謝いたします.

References

[1] W. J. Seiler and M. L. Beall, "Communication: making connections," 2002.

[2] G. Rodman and R. B. Adler, "Style: Delivery and Language Choices," *The New Public Speaker*, 1996.

[3] M. Argyle, F. Alkema, and R. Gilmour, "The communication of friendly and hostile attitudes by verbal and non-verbal signals," *European Journal of Social Psychology*, vol. 1, no. 3, pp. 385–402, 1971.

[4] EnglishClub, "Communicating with Body Language," *Technically Speaking*, p. 2, 2002.

[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded Pyramid Network for Multi-Person Pose Estimation,"

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *Cvpr*, vol. 1, p. 7, 2017.

[7] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional Multi-person Pose Estimation," nov 2016.

[8] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards Accurate Multi-person Pose Estimation in the Wild," jan 2017.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," mar 2017.

[10] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 66–78, 2011.

[11] D. A. Silverstein and Tong Zhang, "System and method of providing evaluation feedback to a speaker while giving a real-time oral presentation," apr 2006.

[12] R. Hincks and J. Edlund, "Promoting increased pitch variation in oral presentations with transient visual feedback," *Language Learning & Technology*, vol. 13, no. 3, pp. 32–50, 2009.

[13] T. Gao, C. Wu, H. A. C. V. W. (ICCV, and undefined 2009, "User-centric speaker report: Ranking-based effectiveness evaluation and feedback," *ieeexplore.ieee.org*.

[14] K. Kurihara, M. Goto, and J. Ogata, "Presentation sensei: a presentation training system using speech and image processing," *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pp. 358–365, 2007.

[15] A. T. Nguyen, W. Chen, and M. Rauterberg, "Intelligent presentation skills trainer analyses body Movement," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9095, pp. 320–332, 2015.

[16] J. Heer, N. Good, A. Ramirez, M. D. .... C. on human ..., and undefined 2004, "Presiding over accidents: system direction of human action," *dl.acm.org*.

[17] M. Goto, K. Itou, T. Akiba, and S. Hayamizu, "Speech Completion: New Speech Interface with On-demand Completion Assistance," *Proceedings of HCI International*, vol. 1, pp. 198–202, 2001.

[18] J. Shotton, A. Fitzgibbon, A. Blake, A. Kipman, M. Finocchio, B. Moore, and T. Sharp, "Real-Time Human Pose Recognition in Parts from a Single Depth Image," jun 2011.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," jun 2014.

[20] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," jan 2016.

[21] Z. Zhang, "Microsoft Kinect Sensor and Its Effect," *IEEE Multimedia*, vol. 19, pp. 4–10, feb 2012.

[22] P. H. Zimmerman, J. E. Bolhuis, A. Willemsen, E. S. Meyer, and L. P. J. J. Noldus, "The Observer XT: A tool for the integration and synchronization of multimodal signals," *Behavior Research Methods*, vol. 41, pp. 731–735, aug 2009.