제25회 한국정보처리학회 춘계학술발표대회 논문집 제13권 제1호 (2006. 5)

비디오 편집을 위한 손동작 추적 및 인식

박호식*, 차승주*, 정하영*, 나상동**, 배철수*
*관동대학교 전자통신공학과
**조선대학교 컴퓨터공학부
e-mail : mediana@netsgo.com

Hand Gesture Tracking and Recognition for Video Editing

Ho-Sik Park*, Seung-Joo Cha*, Ha-Young Jung*, Sang-Dong Ra**, Cheol-Soo Bae*
*Dept. of Electronic Communication Eng, Kwandong University
**Dept. of Computer Eng, Chosun University

요 약

본 논문에서는 동작에 근거한 새로운 비디오 편집 방법을 제안한다. 강의 비디오에서 전자 슬라이드 내용을 자동으로 검출하고 비디오와 동기화한다. 각 동기화된 표제의 동작을 연속적으로 추적 및 인식한 후, 등록된 화면과 슬라이드에서 변환 내용을 찾아 동작이 일어 나는 영역을 확인한다. 인식된 동작과 등록된 지점에서 슬라이드의 정보를 추출하여 슬라이드 영역을 부분적으로 확대한다거나 원본 비디오를 자동으로 편집함으로써 비디오의 질을 향상 시킬 수가 있다. 2 개의 비디오 가지고 실험한 결과 각 95.5, 96.4 %의 동작 인식 결과를 얻을 수 있었다

1. 서론

본 논문에서는 비디오 편집을 목적으로 강의시 동작을 분석하는데 주안점을 두고 있다. 일반적으로 하나의 카메라를 고정하여 강사와 전자슬라이드 화면을 전반적으로 촬영함으로써 움직임이 부족하고 비디오 편집이 어렵다.

기존의 강의 비디오의 자동 편집 시스템은 다수의 카메라로부터 여러 화면을 합쳐 사용[6,9]하고 있다. 이 경우 편집은 미리 정의된 규칙에 의하여 카메라 사이의 화면을 바꾸는 것이다.

본 논문에서는 보다 정확하고 자연스러운 편집을 위하여 규칙에 의한 편집 대신에 동작 인식에 의한 편집 방법을 제안한다. 예를 들어 발표자가 화면의 특정 영역을 지시하거나 원을 그릴 때 편집 된 비디오에서는 그 영역을 확대하여 보여주는 것이다. 그러기 위해서는 먼저 2 가지 문제를 해결해야 된다. 1) 동작을 효과적으로 추적하고 인식하는 문제 2) 저해상도와 불규칙한 조명에서 제공되는 비디오에서 내용 편집을 실행하는 문제이다. 첫 번째 문제는 피부 색상 검출 및 추적에 HMM(hidden Markov model)을 적용하여 해결하였고, 두 번째 문제는 문자 정합으로 이룬 비디오 화면과 전자 슬라이드에서 가져온 지점을 사용하였다. 화면과 슬라이드의 정합이 이루어지면 인식된 동작과 등록된 지점에서 슬라이드의 정보를 추출하여 슬라이드 영역을 부분적으로 확대한다거나 원본 비디오를 자동으로 편집함으로써 비디오의 질을 향상 시킬 수가 있다

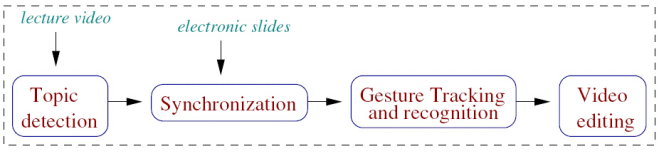


그림 1. 전체 시스템 개요.

그림 1 은 제안된 방법의 전체적인 개요이다. 비디오에서 표제를 초기화하고 검출한다. 검출된 화면은 비디오 문자 분석에 의해 전자 슬라이드와 동기화된다. 각 동기화된 화면을 위해 순차적으로 후보 동작을 추적하고 인식한다. 인식된 동작을 활용하여 비디오 편집에 기준으로 삼는다.

2. 화면과 슬라이드의 동기화

본 논문에서는 하나의 정지된 비디오 카메라로 강의 비디오를 획득하였다. 그림 2 와 같이 발표자는 카메라 앞에서 자유롭게 움직이도록 하였다. 주제 분류를 위하여 감압 없는 DC 비디오 문자 분석[10]을 사용하였고, 동기화를 위해 비디오 문자 검출 방법[12]을 적용하여 고해상도 재건, 이진화 및 문자 인식을 하였다. 이러한 방법은 문자 인식 전에 고해상도 비디오 문자 재건을 활용하여 이루어진다. 또한, 표제와 내용의 유사성을 측정하여 화면과 슬라이드의 정합[12]을 이룬다

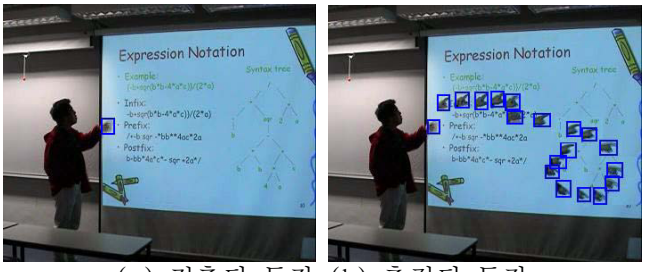
3. 동작 추적

본 논문에서는 투영된 슬라이드의 영역에서 동작을 추적하는데 주안점을 두고 있다. 동기화 이후 프레임간 차이를 이용하여 동작 검출 및 위치를 파악하고 피부 색상 검출한다. 한번 동작이 검출되면 슬라이드 영역 밖으로 나가기 전 까지 추적을 계속해나간다.

3.1 피부 색상 검출

매 10 개의 프레임 중 초기 하나의 프레임을 선택하고 현재와 이전 프레임간의 차이를 계산한다. 슬라이드 영역 안에서 동작의 변화가 나타나면 후보동작을 획득하고, 동작이 위치한 지점을 분석하여 피부 색상을 검출해낸다. 피부 색상 검출은 얼굴 및 동작 인식을 위한 유용하고 견실한 방법이다. 피부 색상 모델링을 위해 색상 차이는 RGB, 표준화된 RGB, HSV, YUV 및 YCrCb 을 이용한다.

가우시안 및 베이즈 방법을 포함한 다양한 모델들이 피부색상 분산을 위해 제안되었었다. 이러한 모델들은 일반적으로 구분을 위해 많은 양의 학습데이터를 필요로 한다. 색상 공간에서의 피부 영역에 대한 명백한 정의에 의한 색상 분류자를 구성하여야 한다. 각 화소는 RGB 색상에서 $\{R > 95, G > 40, B > 20, |R - G| > 15, R > G, R > B\}$ 이거나 $\{R > 220, G > 210, B > 170, |R - G| \leq 15, R > B, G > B\}$ 일 때 피부 색상으로 분류하였다. 이러한 방법으로 분류자를 쉽게 구축할 수 있다. 분류된 피부 색상 화소는 너무 작거나 큰 군집은 제거하였다. 그림 2 (a)에 제안된 방법에 의해 손 동작을 인식한 결과를 나타내었다.



(a) 검출된 동작 (b) 추적된 동작
그림 2. 동작 검출 및 추적.

3.2 피부 색상 추적

동작이 한번 검출되면 매 3 프레임마다 동작을 추적하게 된다. 피부 색상 검출 및 군집화에 의해 추적이 이루어진다. 검출된 피부-색상 화소는 몇 개의 군집으로 그룹화 된다. 추적된 동작은 다음과 같은 조건을 만족하여야 한다. 1) 과거 위치와 근접, 2) 대략적인 동일한 차원, 3) 유사한 색상 분포. 그림 2(b) 에 추적된 동작의 움직임 궤적을 나타내었다.

4. 동작 인식

동작 인식을 위해 이산 HMM 모델을 사용하였다. 본 논문에서는 회전, 줄치기, 지시에 대한 3 가지 동작을 정의하였다.

4.1 동작 분할

손의 움직임 경로는 몇 개의 동작과 일부 동작이 아닌 움직임으로 구성된다. 동작 인식 전에 하나의 동작만을 포함하는 몇 개의 부분으로 나누는 분할된 경로가 필요하다. 본 논문에서는 휴식 지점 검출에 의한 분할 방법을 사용하였다. 휴식 지점은 다음의 조건에 의해 식별 할 수 있다.

1. 전,후 지점에 비하여 급속한 움직임 속도의 변화가 있는 경우.
2. 시작 지점과 매우 근접하거나 이전 지점이 경로상에 있는 경우.
3. 전,후 지점에 비하여 급격하게 움직임 방향이 바뀌는 경우.
4. 시작 지점에서 가장 먼 지점에 있는 경우.

각 움직임 경로는 이러한 조건들을 하나씩 검사한다. 조건 1 이 가장 높은 우선 순위를 가진다. 동작시 손은 부드럽게 움직인다고 가정하였다. 이 조건은 다른 동작들간의 연결 지점과 손이 슬라이드 영역으로 들어오거나 나가는 것을 대다수 찾을 수 있었다. 그리고 조건 1 에 의해 분할된 각 지점을 조건 2 를 만족하는지 확인하였다. 이 목적은 폐원을 이루는지 확인하는 것이다. 두 개의 끝 지점 사이의 경로는 지시 동작을 제외하고는 충분히 큰 높이와 너비를 가져야 하고, 조건 3, 4 도 만족하여야 한다.

측정된 경로는 이러한 휴식지점에서 몇 개의 부분으로 분할된다. 각 부분은 원, 선 지시와 같은 동작이나 도형 혹은 일부 동작이 아닌 동작 중 하나이다. 각 영역은 동작이 아닌 경우를 제외하고는 최소 10 프레임 이상 지속된다. 또한, 처음과 마지막 부분은 슬라이드 영역으로 들어오고 나가는 것으로 간주한다. 모든 분할 된 영역은 원, 선, 지시 그라고 비동작 4 가지로 구분된다. 그림 3 에 이러한 단계를 나타내었다. 그림 3(a)-(c) 는 동작 경로에서의 3 프레임을 나타내고 있다. A, B, C, D 는 조건 1 에 의해, E, F 는 조건 2 로 찾아진 부분이고 IA, BC, ED 그리고 DO 는 분할은 제외된 부분이다.

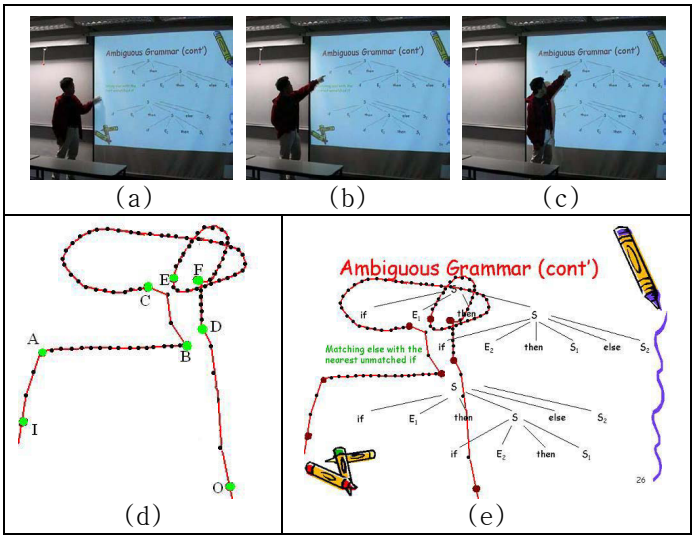


그림 3. 동작 분할

4.2 특징 추출

각 동작 경로 중 20 개의 지점을 특징점으로 선택하였다. v_m을 시작 지점 v_0에서 가장 먼 지점, $D = |v_0-v_m|$ 이라 하자. 각각의 지점 v_i는 두 가지 특징을 계산한다. v_0에서 관계 거리인 $d_i = |v_0-v_i|/D$ 와 벡터 v_{i-1}와 v_0 사이의 각 ϕ_i 이다. d_i 와 ϕ_i는 총 10 단계로 양자화하였다.

4.3 HMM 인식

HMM 은 음성인식에서 뛰어난 성능을 나타내소 있고, 최근에는 언어 번역이나 동작 인식에도 적용되고 있다. 신경망 과 동적시간정합과 같은 다른 방법에 비해 HMM 의 손과 같은 직무에 가장 큰 장점은 선택적이고. 식견 있고, 크기 조절 가능한 재단 모델이라는 것이다. 본 논문에서는 HMM 평가와 추정문제를 해결하기 위해 Viterbi 와 Baum-Welch 알고리즘을 사용하였다. 상태의 수는 줄치기에 대해서는 6 을, 회전 및 지시에 대해서는 8 을 사용하였다.

5. 비디오 편집

동작의 인식으로 교습 내용의 주안점은 알 수 있다. 카메라를 클로즈업하여 슬라이드 영역을 부분적으로 확대한다거나 원본 비디오 편집시 자동으로 그 부분을 편집하는 것이 효과적일 것이다. 편집시 사용 가능한 전자 슬라이드의 정보를 취하여 비디오의 질을 향상 시킬 수가 있었다.

5.1 비디오와 슬라이드 등록

비디오 획득시 슬라이드 영역은 일반적으로 평행하지 않은 화면에 투영된다. 그러므로 비디오에 실제 슬라이드와 왜곡되게 지점을 등록하기 위해 비디오 영상 $p = (x_i, y_i)$, 슬라이드 영상 $\hat{p} = (\hat{x}_i, \hat{y}_i)$ 이라면 호모그라피 H 를 이용하여 해결 할 수 있다.

$$\begin{bmatrix} \hat{x}_i \\ \hat{y}_i \\ 1 \end{bmatrix} \cong \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \quad (1)$$

그러므로,

$$\hat{x}_i(h_{20}x_i + h_{21}y_i + h_{22}) = h_{00}x_i + h_{01}y_i + h_{02} \quad (2)$$

$$\hat{y}_i(h_{20}x_i + h_{21}y_i + h_{22}) = h_{10}x_i + h_{11}y_i + h_{12} \quad (3)$$

n 개의 지점에 대해서는

$$\begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -\hat{x}_1x_1 & -\hat{x}_1y_1 & -\hat{x}_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -\hat{y}_1x_1 & -\hat{y}_1y_1 & -\hat{y}_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_n & y_n & 1 & 0 & 0 & 0 & -\hat{x}_nx_n & -\hat{x}_ny_n & -\hat{x}_n \\ 0 & 0 & 0 & x_n & y_n & 1 & -\hat{y}_nx_n & -\hat{y}_ny_n & -\hat{y}_n \end{bmatrix} \times \begin{bmatrix} h_{00} \\ h_{01} \\ h_{02} \\ h_{10} \\ h_{11} \\ h_{12} \\ h_{20} \\ h_{21} \\ h_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

이것은 식(4)와 같이 나타낼수 있다.

$$Ah = 0 \quad (4)$$

이것은 선형 최소 제곱 이므로, 식 (5)와 같이 간략하게 할 수 있다.

$$\| Ah \|^2 = (Ah)^T Ah = h^T A^T Ah \quad (5)$$

여기서 h 는 최소 고유값을 가지는 A^TA 의 고유벡터이다. h 는 크기에 대해서만 정의되어지므로 단위벡터 h로 사용하여 해결하였다. 수식(5)는 4 개 이상의 지점을 필요로 한다.

화면과 슬라이드 사이에 상응하는 지점은 인식된 문자의 위치를 활용하여 찾는다. 인식된 비디오 문자의 지점은 전자 슬라이드로부터 추출된 문자의 내용과 지점으로 정합되고, 동기화 되어 확인 할 수 있다. 왜냐하면 표제어 인식은 대체적으로 정확하고 슬라이드 등록을 위하여 표제어를 활용한다. 그림 3(e) 는 등록 이후 슬라이드에 추적된 동작 움직임 경로를 겹쳐 놓은 것이다.

5.2 편집

등록 이후 슬라이드의 정보는 추출되어 보다 나은 시각 효과를 위하여 저해상도의 편집에 추가된다. 그림 4 는 그림 3 의 3 프레임의 편집 결과를 나타낸다. 등록 지점에 의하여 동작이 인식되었을 때 대응하는

슬라이드의 위치를 알 수 있으므로 강사가 주안점을 두고 있는 영역을 클로즈업하여 슬라이드 영역을 부분적으로 확대하게 자동 편집이 가능 할 것이다.

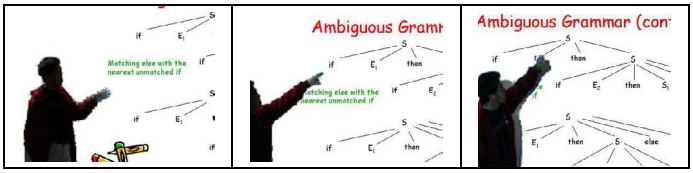


그림 4. 그림 3(a)-(c)의 편집된 비디오.

6. 실험결과 및 고찰

2 개의 비디오를 가지고 실험을 수행하였다. 각 비디오의 지속시간은 40 분이었고, 각 약 30 장의 슬라이드로 구성되어있다. 표 1 은 화면과 슬라이드의 동기화 결과를 나타내고 있다. 두 개의 슬라이드가 같은 표제를 가지고 있고 내용상 충분한 문자를 인식하지 못한 일부의 경우 정합을 이루지 못하였다.

표 1. 동기화 결과

Lecture Video	Total number of slides	# of correctly matched slides	Accuracy
1	34	31	91.2%
2	26	25	96.2%

두 개의 비디오는 총 59,523 개의 프레임으로 구성되었다. 모든 동작들은 바르게 검출되었으며 추적 알고리즘은 동작을 찾지 못한 11 프레임과 잘못 인식한 4 프레임을 제외하고는 전반적으로 바르게 동작하였다. 동작을 찾지 못하거나 잘못 인식한 주된 이유는 투영된 화면에 조명이 부족하거나 많은 경우와 중첩이 되는 경우였다. 동작 인식을 위하여 각 동작에 대해 200 개의 견본을 가지고 HMM 학습을 하였다. 그러나 이것은 학습과 실험 모두에서 충분하지 않은 데이터임으로 약간의 타원이나 선 같은 그림을 비디오로부터 추출된 동작 경로 대신에 학습 데이터로 사용하였다. 실험에서 이러한 지시 학습 데이터는 비디오로부터 실제 동작을 인식하기 위해 사용되었다. 표 2 에 동작 인식 결과를 나타내었다.

표 2. 동작 인식 결과

Video	Gesture	Circling	Lining	Pointing
1	Number of gestures	45	51	105
	Correctly recognized	45	49	98
	Recognition rate	1.00	0.96	0.94
	Overall Recognition Rate	0.955		
2	Number of gestures	35	67	92
	Correctly recognized	34	65	88
	Recognition rate	0.97	0.97	0.96
	Overall Recognition Rate	0.964		

7. 결론

본 논문에서는 동작 검출, 추적 그리고 인식을 기반으로 한 강의 비디오 편집 방법을 제안하였다. 강의 비디오에서 전자 슬라이드 내용을 자동으로 검출하고 비디오와 동기화하고, 각 동기화된 표제의 동작을 연속적으로 추적 및 인식하여 등록된 지점에서 슬라이드의 정보를 추출하여 비디오를 자동으로 편집함으로써 비디오의 질을 향상 시킬 수가 있도록 하였다. 2 개의 비디오 가지고 실험한 결과 각 95.5, 96.4%의 동작 인식 결과를 얻을 수 있었다.

현재는 교실에서 강사의 행동을 추정하는데 손동작만을 이용하였지만, 향후에는 머리 자세와 같은 다른 정보도 추가로 손동작에 비디오 편집을 위한 기능을 향상 시킬 것이다.

참고문헌

[1] G. D. Abowd et. al., "Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project," ACM Multimedia, pp. 187-198, 2000.

[2] S. G. Deshpande & J.-N. Hwang, "A Real-time Interactive Virtual Classroom Multimedia Distance Learning System," IEEE Trans on Multimedia, vol. 3, no. 4, pp. 432-444, 2001.

[3] D. Phung, S. Venkatesh & C. Dorai, "High Level Segmentation of Instructional Videos Based on Content Density," ACM Multimedia, 2002.

[4] L. He, E. Sanocki, A. Gupta & J. Grudin, "Auto-Summarization of Audio-Video Presentations," ACM Multimedia, pp. 489-498, 1999.

[5] S. X. Ju et. al, "Summarization of Videotaped Presentations: Automatic Analysis of Motion and Gesture," IEEE Trans on CSVT, vol. 8, no. 5, pp. 686-696, 1998.

[6] Q. Liu, Y. Rui, A. Gupta & J. J. Cadiz, "Automatic Camera Management for Lecture Room Environment," Int. Conf. on Human Fectirs in Computing Systems, 2001.

[7] T. F. S. -Mahmood, "Indexing for topics in videos using foils," Int. Conf. CVPR, pp. 312-319, 2000.

[8] J. Martin & J. B. Durand, "Automatic Gestures Recognition Using Hidden Markov Models," Int. Conf. Automatic Face and Gesture Recognition, 2000.

[9] S. Mukhopadhyay & B. Smith, "Passive Capture and Structuring of Lectures," ACM Multimedia, 1999.

[10] C. W. Ngo, T. C. Pong & T. S. Huang, "Detection of Slide Transition for Topic Indexing," Int. Conf. on Multimedia Expo, 2002.

[11] P. Peer, J. Kovac & F. SOLINA, "Human skin colour clustering for face detection," Int. Conf. on Computer as a Tool, 2003.

[12] F.Wang, C. W. Ngo & T. C. Pong, "Synchronization of Lecture Videos and Electronic Slides by Video Text Analysis," ACM Multimedia Conference, 2003.