Optical Flow를 이용한 단모음(아,에,이,오,우) 분석

이미애†, 박기수††
†국립한밭대학교 BK21사업단
††고신대학교 컴퓨터과학부

Vowels(a,e,i,o,u) Analysis Using Optical Flow

Mi-Ae Lee†, Ki-Soo Park††
† Center of BK21 , HanBat National University
††Department of Computer Science, Kosin University
E-mail :†malee@hanbat.ac.kr,††pkisoo@kosin.ac.kr

요 약

컴퓨터를 이용한 독순 연구는 Man Machine Interface, 지적부호화에 있어서의 송신측 기술, 청각장애인의 독순 훈련 시스템 등 다방면에서 그 응용이 기대된다. 본 논문은, 움직임 정보는 입술의 에지영역에 집중하고 있음에 주목하여, 입술 에지영역의 Optical Flow 추정값을 독순정보로 이용하는 방법을 제안한다. 휘도값을 갖지 않는 에지에, 선형 가상 휘도값를 정해주어 Optical Flow를 추정하는 VGM을 도입해 특징 파라미터를 계산하고, 마할라노비스 평방거리(Mahalanobis's square distance)에 기초한 최대우도판별함수를 이용하여 단모음을 분석하는 알고리즘을 제안한다.

1. 서론

최근, 컴퓨터의 기술진보와 함께 컴퓨터와 이용자 상호간의 커뮤니케이션 수단으로서, 음성정보, 영상정보를 활용하는 연구가 활발히 진행되고 있다[1]. 그 중, 독순(Lip-reading)연구는 Man Machine Interface, 지적부호화에 있어서의 송신측 기술, 청각장애인의 독순 훈련 시스템[2] 등 다방면에서 그 응용이 기대된다. 독순은 사람의 입술모양, 턱운동, 얼굴표정 등의 시각정보를 통해 음성을 인식하는 방법이다.

종래의 대표적인 독순연구는, 입술형상의 관측으로 얻어지는 특징량을 이용하는 방법[3]과 입술 주위의 움직임벡터인 Optical Flow를 이용하는 방법[4]으로 대별된다. 전자(前者)는 입술의 횡폭(橫幅)과 종폭(縱幅), 측면얼굴의 굴곡정도, 깊이 등을 특징량으로 구해 그 조합으로 단모음(a,e,i,o,u)을 분석했다. 그러나 입술의 특징점을 수작업으로 정한 다음, 두 대의 카메라로 특징량을 구하고 있어, 피험자와의 직접적인 접촉이 불가피하며, 이에 따른 자동인식 시스템구현에 어려움이 있다. 또한, 모음발성 상태의 정지영상만을 이용하므로 실시간처리에의 대응이 곤란하다.

후자(候者)는 독순연구를 발성학의 원리에 기초한 방법으로, 생리학적으로 보면 같은 단어의 발성에는 화자가 바뀌어도 입술주위의 같은 근육이 움직임에 착안하여, 입술주위의 움직임 정보인 Optical Flow 추정값을 패턴화하여, 등록된 단어사전과의 매칭에 의해 분석하였다. 피험자와의 직접적인 접촉없이 영상만으로 처리할 수 있어 저항감을 줄일 수 있고, Gradient Method로 Optical Flow를 구하고 있어 수치계산만으로 비교적 간단히 독순시스템을 구현하고 있다. 그러나 이 방법은 영상 휘도값을 이용하기 때문에 영상획득에 있어 조명에 민감하며, 잡음의 영향을 받기 쉬운 단점이 있다.

본 논문은, 시계열 입술영상에 있어 비교적 단순한 휘도값으로 나누어지는 입술영상의 특성과 움직임 정보는 입술의 에지영역에 집중하고 있음에 주목하여, 입술 에지영역의 Optical Flow 추정값을 독순정보로 이용하는 방법을 제안한다. 필자는 물체의 움직임이 극소인 경우, 에지부에 선형 가상농도치를 정해줌으로

서 정밀도 높은 Optical Flow를 추정하는 방법(Virtual Gradient Method, 이하 VGM으로 표기)을 제안하였다[5]. 본 논문은 VGM에 의한 Optical Flow 추정치를 특정 파라미터로 계산하고, 마할라노비스 평방거리(Mahalanobis's square distance)에 기초한 최대우도판별함수를 이용하여 단모음을 분석하는 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 에지동영상의 Optical Flow 추정방법인 VGM에 대해 설명하고 3장에서 모음분석 방법을 제시한다. 4장에서는 실험을 통해 제안 방법에 대한 유효성을 보이고, 5장의 결론에서 향후의 연구과제를 제시한다.

2. Virtual Gradient Method

본 장에서는 물체의 움직임 정보가 집중되어 있는 에지의 Optical Flow 추정방법으로, Gradient Method의 구속식을 기본식으로 이용하는 VGM에 대해 소개한다. 먼저, 시계열 영상에서 운동물체의 어느 한 점 (x,y)의 시각 t 에 대한 휘도를 $V(x,y,t)$라 하고, 극소시간 δt 후의 대응점의 휘도는 극소시간에 있어 불변하다고 가정할 때, 물체의 움직임벡터인 Optical Flow를 (u,v)라 하면, 다음과 같은 Gradient Method 구속식(1)이 도출된다.

$$V_x u + V_y v + V_t = 0 \qquad (1)$$

여기서, V_x, V_y, V_t 는 영상의 휘도 변화량으로, 영상데이터로부터 직접 구할 수 있다.

시계열 에지영상에서 물체의 움직임이 극소이고, 형상의 변화도 극소임을 전제할 때, VGM은 식(1)를 기본식으로 이용한다. 휘도값을 가지지 않는 에지영상으로부터 V_x, V_y, V_t값을 구하기 위해 에지 패턴간의 기하학적 관계를 이용한다. 에지영상의 휘도 기울기는 에지 패턴의 형상에 의존하고 있으므로, 시각 t에서의 (x,y)위치에 있는 에지 패턴 상(上)의 한 점 P의 가상휘도 기울기는 점 P의 접선에 대해 법선방향 부근에 존재한다고 가정할 때, 점 P에 관한 접선각을 이용할 수 있다. 이것을 그림1에 나타내었다. θ는 점 P에 대한 접선각, d는 점 P와 시각 $t+\delta t$에 있어서의 에지패턴 상의 가장 가까운 점과의 거리, φ는 P와 시각 $t+\delta t$에서의 에지 패턴 상의 가장 가까운 점 P'을 연결하는 선과 그 점의 법선과의 사이에 이루어지는 각도이다. 이러한 관계에 의해 V_x, V_y, V_t는 다음과 같이 유도된다.

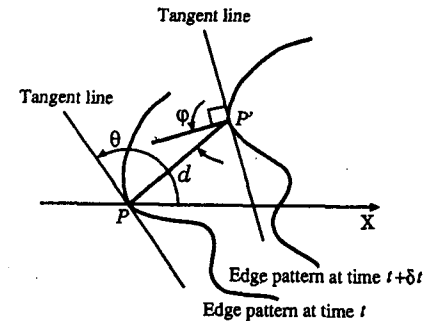


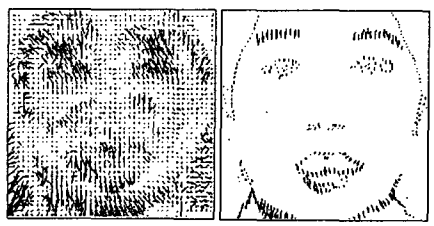
그림1 Virtual Gradient Method의 파라미터 설정

$$\begin{aligned} V_x &= C \cdot \sin\theta \\ V_y &= C \cdot \cos\theta \\ V_t &= d \cdot \cos\varphi \end{aligned} \qquad (2)$$

이 때, C는 점 P의 접선 방향에 대한 최대 휘도 기울기값이며, 접선각 θ는 Chain Code를 이용하여 구하고 있다. V_x, V_y, V_t의 값을 구속식(1)에 대응시킴으로서 에지부에 대한 Optical Flow를 추정할 수 있다.

그러나, 구속식(1)만으로는 각 점의 Optical Flow (u,v)를 한번에 결정할 수 없다. Horn 등은 영상 공간에서의 속도분포는 smoothing하게 변한다는 제약조건을 부가해, 이 제약조건을 어느 정도 만족하는가를

(a) 시계열 얼굴영상



(b) Gradient Method (c) VGM

그림2 Optical Flow 결과

나타내는 평가함수를 구성함으로서, 최소화문제로 풀고 있다[6]. 본 논문에서는, 국소영역 내의 Optical Flow는 같은 값을 가진다고 가정하고 최소2승법으로 (u, v)를 구하였다. 시계열 얼굴영상을 이용하여 Gradient Method와 VGM의 Optical Flow 추정결과를 그림2에 나타내었다.

3. 모음분석 방법

3.1 특징파라미터 추정

모음 분석을 위한 입술영상 취득에는, 얼굴영상으로부터 얼굴의 구조적 특성을 이용하는 방법, 얼굴영상의 칼라값을 이용하는 방법 등 다양한 수법이 제안되고 있다. 본 논문은 입술영상을 취득하기 위한 이러한 전처리과정을 간단히 하기 위해 배경이 없는 시계열 정면얼굴영상의 에지히스토그램을 이용해 입술영역을 구한다. 그리고, 발성에 의한 입술형상 변화의 기본이 되는 단모음 중 <아,에,이,오,우>에 대하여 분석한다.

먼저, 입술영상의 움직임 영역을 그림3처럼 설정한다. 종축(縱軸)방향의 움직임벡터 u에 대해 S_{left}와 S_{right} 영역, 횡축(橫軸)방향의 움직임벡터 v에 대해 S_{upper}와 S_{lower}영역을 설정하고, 시계열 입술영상의 에지영역에 따라 Optical Flow를 추정한다. 본 논문에서는 모음분석을 위한 특징파라미터로 S_{left}영역의 Optical Flow 추정값 u와 S_{lower}영역의 Optical Flow 추정값 v를 이용한다. 연속 두 프레임 i 에 대한 Optical Flow u와 v는 설정영역 S_{left}와 S_{lower}의 평균 Flow로 구할 수 있다.

$$u_i = \frac{1}{Flow\text{의 갯수}} \sum_{S_{left}} u \tag{3}$$

$$v_i = \frac{1}{Flow\text{의 갯수}} \sum_{S_{lower}} v \tag{4}$$

여기서, Flow의 갯수란 설정영역 안에서 추정된 Optical Flow의 총갯수이다. 시계열 입술 에지영상의

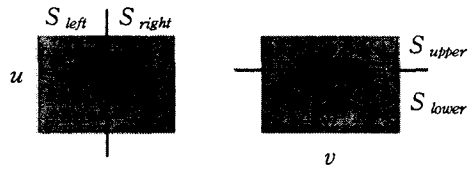


그림3 입술의 특징영역

설정영역 S_{left}와 S_{lower}의 평균 Optical Flow를 식(3), (4)를 이용해 구한 뒤, 각 모음 발성에 의한 입술표현이 최상으로 이루어진 상태의 프레임 수를 n이라 할 때, u_i, v_i의 적분값을 계산한다(식(5),(6)).

$$u = \sum_{1}^{n-1} u_i \tag{5}$$

$$v = \sum_{1}^{n-1} v_i \tag{6}$$

이 때, 연속 프레임에서의 n의 결정방법은, u_i, v_i Flow의 부호가 바뀌는 시점의 프레임을 기준으로 한다. 식(5)에 의해 구해진 u를 횡축 특징파라미터로, 식(6)에 의해 구해진 v를 종축 특징파라미터로 하는 분포도를 이용하여 모음분석을 행한다.

3.2 분석방법

피험자의 모음발성에 의해 얻어진 시계열 입술영상으로부터 모음을 분석하기 위한 방법으로, 마할라노비스 평방거리를 도입한다.

$$D_p^2 = (X-\mu)^t \Sigma^{-1} (X-\mu) \tag{7}$$

여기서, X는 변수벡터, μ는 평균벡터, Σ^{-1}은 분산공분산행렬의 역행렬로 구할 수 있다. D는 P차원에서 중심 μ로부터의 주성분 축상(軸上)의 거리를 나타내며, 자유도 P의 χ^2분포에 따른다. 본 논문에서는 마할라노비스 평방거리 D_2^2 가 자유도2, 유의수준 5%의 χ^2값: $\chi^2(2;0.05)$의 등거리타원으로, 그 결과를 분포도로 나타낸다.

먼저, 피험자들의 각 모음입술영상으로부터 추정된 각각의 특징 파라미터를 이용하여 마할라노비스 평방거리의 등거리타원 표준판별 샘플을 구한다. 어떤 피험자가 모음 k를 발성했을 때, 추정된 특징파라미터 u, v 의 각 카테고리가 $N(\mu_k, \Sigma_k)$의 다변량정규분포에 따른다고 가정하면, 그 밀도함수는 다음과 같다.

$$f(X) = (2\pi)^{\frac{-p}{2}} |\Sigma_k|^{\frac{1}{2}} \times \exp\left\{ -\frac{1}{2} (X-\mu_k)^t \Sigma_k^{-1} (X-\mu_k) \right\} \tag{8}$$

이것으로, 다음의 최대우도판별함수(9)를 구할 수 있다.

$$S_k(X) = -\log|\Sigma_k| - (X-\mu_k)^t \Sigma_k^{-1}(X-\mu_k) \quad (9)$$

미지의 X_i가 주어질 때, $S_k(X_i)$의 값을 최대로 하는 카테고리에 X_i가 속한다고 판단할 수 있다.

4. 실험

시계열 얼굴영상은, 조명의 변화가 크게 없는 실내에서 CCD카메라(33매/sec)로 촬영하였다. 얼굴은 정면방향으로, 머리는 움직이지 않게 고정했다. 4명의 피험자(남자2명,여자2명)를 대상으로 단모음<아,에,이,오,우>을 1set로, 각 50set씩 준비했다. 입술영상은 모음발성을 시작한 첫 얼굴영상의 에지히스토그램을 이용해 입부분을 탐색한 다음, 얼굴의 구조적 위치 정보를 이용하여 입술영역을 정했다(그림4). 입술영상의 에지를 얻기위해 GL filter 처리를 한 후, 에지의 윤곽부분만 추출하여 Optical Flow를 추정했다.

식(3),(4)와 식(5)(6)을 이용해 구해진 피험자들의 단모음 50set의 특징파라미터 중, 5set씩, 총 20set를 이용해 마할라노비스 평방거리에 의한 둥거리타원 모음 표준판별 샘플을 작성하고 나머지 45set는 모음분석 데이터로 이용하였다. 한 피험자의 결과를 그림5에 나타냈다. 각 모음발성에 의한 특징 파라미터들이 소수의 차이는 보이지만 표준판별 샘플을 중심으로 같은 카테고리를 형성하고 있음을 확인할 수 있었다. 모음<아>에 대해선 약 90%이상, <오,우>는 약 80%이상, <에,이>는 70%이상의 인식률을 얻었다.

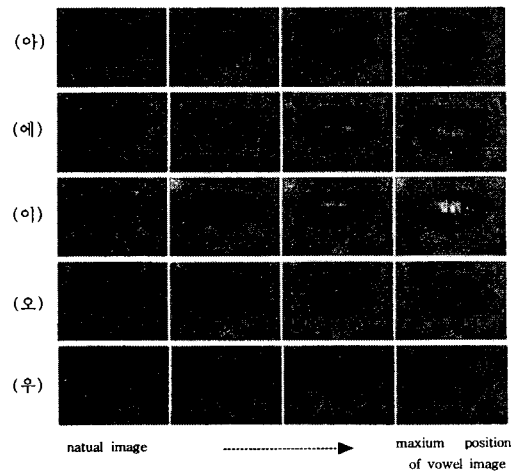


그림4 시계열 입술영상

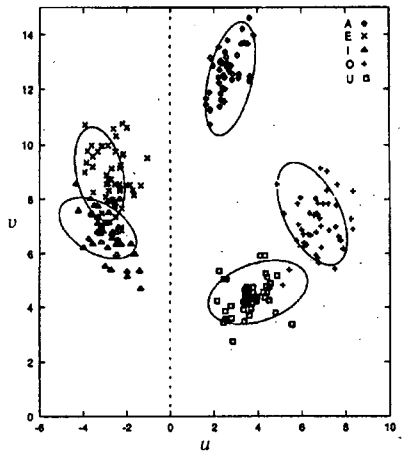


그림5 표준판별 샘플에 의한 모음분석

5. 결론

본 논문은 시계열 입술에지영상의 Optical Flow 추정값을 특징파라미터로 하여, 단모음을 분석하는 방법에 대해 제안했다. 실험을 통해 단모음 모두 70%이상의 인식률을 가짐을 확인하였다. 비록, 판별샘플에는 속하지 않더라도, 같은 모음은 같은 카테고리에 근접해 있음을 알 수 있었다. 향후, 인식률 향상 가능성이 있는 이러한 특징파라미터에 대해서도 방법을 보완할 예정이다.

[참고문헌]

[1] H. Kobayash, F. Hara, "A Basic Study of Man-Machine Interactive Communication by Using Facial Expressions," Human Interface, Oct. pp.18-20, 1993

[2] K. Matsuoka, K. Kurosu, "Speechreading Trainer for the Hearing-Impaired Children," IEICE, Vol. J70-D, No.11, pp.2167-2171, 1987

[3] E. K. Finn, A. A. Montgomery, "Automatic Optically-Based Recognition of Speech," Pattern Recognition Letters, 8, 3, pp.159-164, 1988

[4] K. Mase, A. Pentland, "Automatic Lipreading by Optical-Flow Analysis," IEICE, Vol.J73-D-II, No.6, pp.796-803, 1990

[5] M. Lee, T. Ito, Y. Kaneda, "Motion Extraction of Time Varying Images Using Virtual Gradient Method," ISCIE, Vol.11, No.9, pp.483-490, 1998

[6] B. K. P. Horn and B. G. Schunck, "Determining Optical Flow," Artificial Intelligence, Vol.17, pp.185-203, 1981