Check for updates

Journal of KIIT. Vol. 24, No. 10, pp. 235-245, Oct. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.10.235

비강체 혀 움직임 분석을 위한 신뢰도 기반 3-프레임 광류 보정에 관한 연구

김 근 호*

A Study on the Confidence-Aware Three-Frame Optical Flow Correction for Non-Rigid Tongue Motion

Keun Ho Kim*

이 연구는 한국한의학연구원의 전략연구단사업과 기본사업의 지원을 받았음(과제 번호: SSN2611011, KSN2512012)

요 약

혀는 표면의 색상과 질감이 유사하고 위치에 따라 움직임이 달라지는 비강체 구조로, 픽셀 수준의 밀집 기준 광류(GT, Ground Truth)를 구축하기 어렵다. 본 연구에서는 사전학습된 RAFT와 GMFlow를 이용하여 2-프레임 기준 방법(2F), 시간적 잔차를 직접 보정하는 Direct 3F, 신뢰도에 기반하여 보정을 선택적으로 적용하는 Adaptive 3F를 비교하였다. 네 개의 3-프레임 혀 영상 세트에서 9개 해부학적 기준점을 평가한 결과, Things 조건에서 RAFT의 기준점 오차는 각각 4.92, 6.24, 4.91 픽셀, GMFlow는 8.30, 8.85, 8.24 픽셀이었다. Sintel 조건에서도 Direct 3F의 오차는 증가한 반면 Adaptive 3F는 2F와 유사한 수준을 유지하였다. 이는 시간적 일관성의 향상이 해부학적 추적 정확도의 향상을 보장하지 않으며, 비강체 혀 영상에서는 신뢰도 기반의 선택적 시간 보정이 필요함을 시사한다.

Abstract

The tongue is a non-rigid structure with similar surface color and texture and spatially varying motion, making dense optical-flow Ground Truth (GT) difficult to obtain. This study compared a two-frame baseline (2F), Direct 3F, and confidence-based Adaptive 3F using pretrained RAFT and GMFlow models. Nine anatomical landmarks were evaluated in four tongue image triplets. Under the Things condition, landmark errors for 2F, Direct 3F, and Adaptive 3F were 4.92, 6.24, and 4.91 pixels for RAFT and 8.30, 8.85, and 8.24 pixels for GMFlow, respectively. Under the Sintel condition, Direct 3F increased landmark error, whereas Adaptive 3F maintained accuracy comparable to 2F. These results show that improved temporal consistency does not necessarily improve anatomical tracking accuracy and support confidence-based selective temporal correction for non-rigid tongue motion.

Keywords

optical flow, tongue motion, non-rigid motion, temporal consistency, anatomical landmark, pretrained model, RAFT, GMFlow

* 한국한의학연구원 디지털임상연구부
- ORCID: https://orcid.org/0000-0001-6688-6674

· Received: Sep. 18, 2026, Revised: Oct. 02, 2026, Accepted: Oct. 05, 2026

· Corresponding Author: Keun Ho Kim
Digital Health Research Division
Korea Institute of Oriental Medicine
Tel.: +82-42-868-9365, Email: rkim70@kiom.re.kr

Ⅰ. 서 론

혀는 발음, 저작 및 연하 과정에서 복잡하고 정교한 움직임을 수행하며, 이러한 움직임은 혀의 근육 및 신경학적 조절과 밀접하게 관련되어 있다[1][2]. 따라서 연속 영상에서 혀의 움직임을 정량적으로 추적하는 기술은 혀의 운동 특성을 객관화하고 비정상적인 움직임을 분석하기 위한 기초 기술로 활용될 수 있다.

영상 기반 움직임 분석은 연속 영상에서 객체의 위치 및 형태 변화를 추적하는 데 활용되어 왔으며, 광류를 이용한 객체 위치 추적과 구강 영역의 움직임 분석 등 다양한 분야에 적용되고 있다[3][4]. 광류(Optical flow)는 연속 영상 사이의 픽셀 변위를 추정하여 움직임의 방향과 크기를 밀집 벡터장(Dense vector field)로 표현하는 방법으로, 전통적인 Farnebäck 방법에서부터 최근의 딥러닝 기반 RAFT(Recurrent All-Pairs Field Transforms)와 GMFlow(Global Matching Optical Flow)에 이르기까지 다양한 방법이 제안되어 왔다[5]-[7]. 그러나 혀 영상은 일반적인 자연영상과 비교하여 광류 추정에 불리한 특성을 갖는다. 혀 표면은 국소적으로 유사한 색상과 질감이 나타나며, 타액에 의한 표면 반사와 촬영 조건에 따른 밝기 변화가 발생할 수 있다. 특히 혀 표면의 수분은 영상에서 광택 또는 정반사로 나타날 수 있어 안정적인 영상 특성 추출에 영향을 줄 수 있다[8]. 이러한 특성은 연속 프레임 사이에서 안정적인 픽셀 대응을 추정하는 데 어려움을 줄 수 있다.

또한 혀는 강체가 아닌 비강체 구조로서 위치에 따라 서로 다른 변형과 움직임이 발생할 수 있다. 혀의 순간적인 형태는 내재근 및 외재근의 복합적인 작용에 의해 변화하므로[1], 하나의 전역적인 움직임만으로 전체 혀의 운동을 설명하기 어렵다. 따라서 광도 기반 영상 재구성이나 광류의 자기 일관성이 향상되더라도 실제 해부학적 위치의 추적 정확도가 함께 향상된다고 단정하기 어렵다.

한편 혀 영상에 특화된 광류 네트워크를 지도학습하기 위해서는 연속 프레임 사이의 실제 픽셀 변위를 나타내는 픽셀 수준의 밀집 기준 광류(GT, Ground Truth)가 필요하다. 그러나 실제 혀 표면의 모든 픽셀에 대한 변위를 직접 측정하여 밀집 기준 광류를 구축하는 것은 현실적으로 어렵다. 이러한 밀집 기준 광류 확보의 한계는 광류 연구에서 합성 데이터셋이 활용되는 중요한 배경 중 하나이며, FlyingThings3D와 같은 데이터셋은 광류 및 장면 흐름(Scene flow) 학습을 위한 대규모 합성 기준 데이터를 제공한다[9]. 따라서 본 연구에서는 제한된 혀 영상으로 광류 네트워크를 새롭게 학습하지 않고, 기존 대규모 데이터로 사전학습된 RAFT와 GMFlow를 이용하여 밀집 광류를 추정하였다.

그러나 자연영상 또는 합성 데이터에서 학습된 광류 모델이 혀의 비강체 움직임에 동일하게 일반화된다고 가정하기는 어렵다. 이에 본 연구에서는 밀집 기준 광류 대신 연속 영상에서 수동으로 지정한 9개의 해부학적 기준점(Anatomical landmark)을 독립적인 평가 기준으로 사용하여 광류에 의한 해부학적 위치 추적 정확도를 평가하였다. 또한 주 실험에서는 Things 계열 체크포인트(Checkpoint)를 사용하고, 사전학습 조건에 따른 결과의 민감도를 확인하기 위해 MPI-Sintel 관련 체크포인트를 이용한 보조 실험을 수행하였다[9][10].

기존 RAFT와 GMFlow는 기본적으로 두 영상 사이의 광류를 추정하므로, 세 번째 프레임의 정보를 이용하면 추가적인 시간적 일관성을 평가할 수 있다. 연속된 세 프레임에서 1→2와 2→3 광류를 합성한 결과와 1→3에서 직접 추정한 광류 사이의 차이를 시간 잔차로 정의할 수 있다. 그러나 이 시간 잔차에는 F_{12}뿐만 아니라 F_{23}과 F_{13}의 추정오차도 함께 포함될 수 있으므로, 시간 잔차를 직접 F_{12}의 보정에 사용하는 것이 실제 해부학적 대응을 향상시키는지는 별도로 검증할 필요가 있다.

이에 본 연구에서는 2-프레임 기준 방법(2F), 시간 잔차를 직접 적용하는 Direct 3F, 그리고 시간적 일관성, 순방향-역방향 일관성 및 방향 일관성(Direction consistency)을 기반으로 보정을 선택적으로 적용하는 Adaptive 3F를 비교하였다. 영상 재구성 오차(WE, Warping Error), 순방향-역방향 일관성 오차(FBE, Forward-Backward consistency Error), 시간 오차(TE, Temporal Error) 및 9개 해부학적 기준점의 기준점 오차(LE, Landmark Error)를 평가하여, 3-프레임 시간적 일관성의 향상이 실제 해부학적 추적

정확도의 향상으로 이어지는지를 분석하고 비강체 혀 움직임에 대한 선택적 시간 보정의 유효성을 평가하고자 하였다.

II. 연구 방법

2.1 혀 영상 및 해부학적 기준점

본 연구에서는 시간적으로 연속된 혀 영상 가운데 혀 영역이 배경 영상으로부터 분할된 T1, T2, T3 및 T4의 네 개의 3-프레임 영상 세트를 사용하였다. T1은 상하 방향의 변위와 형태 변형이 모두 작은 경우, T2는 상하 방향의 변위가 상대적으로 큰 경우, T3는 좌우 방향의 움직임이 두드러진 경우, T4는 형태 변형이 상대적으로 큰 경우로 선정하였다. 원본 혀 영상의 해상도는 가로 480 × 세로 660 픽셀이며, 초당 10 프레임(FPS, Frame Per Second)으로 촬영하였다. 각 영상 세트는 세 프레임 I1, I2 및 I3로 구성하였으며, 모든 비교 방법에서 동일한 영상과 전처리 조건을 사용하였다.

광류 벤치마크에서는 모든 픽셀의 실제 변위를 나타내는 밀집 기준 광류를 이용하여 추정 광류의 정확도를 평가할 수 있다[10]. 그러나 실제 혀 영상에서는 이러한 픽셀 수준의 밀집 기준 광류를 직접 구축하기 어렵기 때문에, 본 연구에서는 해부학적 추적의 독립적인 평가 기준으로 각 프레임에 9개의 기준점을 수동 지정하였다. 기준점은 좌상(1), 좌(2), 좌하(3), 설근(4), 중앙(5), 설첨(6), 우상(7), 우(8) 및 우하(9)로 그림 1과 같이 구성하였다. 수동 기준점 지정의 관찰자 내 반복성을 평가하기 위하여 9개 기준점 중 혀의 주요 해부학적 위치를 대표하는 5개 기준점(설근, 중앙, 설첨, 좌측 및 우측)을 선정하였다. 동일 평가자가 각 프레임에서 이들 5개 기준점을 5회 반복하여 지정하였으며, 각 기준점의 반복 측정 좌표와 5회 평균 좌표 사이의 유클리드 거리를 기준점 지정 오차(Annotation error)로 정의하였다.

첫 번째 프레임의 i번째 기준점 $P_{1,i}$에서 추정된 광류 F_{12}를 이중선형 보간(Bilinear interpolation)하여 두 번째 프레임에서의 예측 위치 $\hat{P}_{2,i}$를 식 (1)과 같이 계산하였다.

$$\hat{P}_{2,i} = P_{1,i} + F_{12}(P_{1,i}) \tag{1}$$

예측 위치와 두 번째 프레임에서 수동 지정한 i번째 기준점 $P_{2,i}$ 사이의 유클리드 거리를 LE로 식 (2)와 같이 정의하였다.

$$LE_i = \| \hat{P}_{2,i} - P_{2,i} \|_2 \tag{2}$$

각 영상 쌍의 기준점 오차는 9개 기준점에서 계산한 LE의 평균값으로 산출하였다.

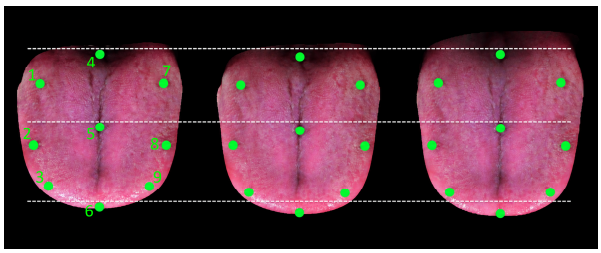


그림 1. 세 연속 혀 영상과 9개 해부학적 기준점의 예
Fig. 1. Example of three consecutive tongue images and nine anatomical landmarks

2.2 사전학습 광류 모델 RAFT와 GMFlow

실제 혀 표면의 모든 픽셀에 대한 밀집 기준 광류를 확보하기 어렵기 때문에 본 연구에서는 혀 영상만을 이용한 별도의 지도학습을 수행하지 않았다. 대신 기존 대규모 광류 데이터로 사전학습된 모델을 이용하였으며, 서로 다른 광류 추정 구조를 비교하기 위하여 RAFT[6]와 GMFlow[7]를 비교 모델로 선정하였다. RAFT는 모든 쌍 상관 볼륨(All-pairs correlation volume)을 구성하고 순환 갱신(Recurrent update)을 통해 광류를 반복적으로 개선하는 구조로, 국소적인 움직임뿐 아니라 비교적 큰 변위에 대해서도 대응 관계를 반복적으로 정제할 수 있다는 장점이 있으나, 모든 쌍 상관 정보를 구성하고 반복적으로 갱신하므로 계산량과 메모리 요구가 증가할 수 있다[6]. 반면 GMFlow는 광류 추정을 전역 정합(Global matching) 문제로 구성하여 특징 대응(Feature correspondence)을 직접 추정하므로, 장거리 대응을 명시적으로 모델링할 수 있다는 장점이 있으나, 영상의 특징 표현과 전역 정합의 품질에 따라 대응

정확도가 영향을 받을 수 있다[7]. 이처럼 두 모델은 각각 반복적 순환 개선과 전역 정합이라는 서로 다른 광류 추정 방식을 사용하므로, 본 연구에서는 시간 보정 방법의 효과가 광류 추정 구조에 따라 달라지는지를 비교하기 위한 모델로 선정하였다.

주 실험에서는 FlyingThings3D를 포함한 Things 계열의 대규모 합성 광류 데이터로 학습된 체크포인트를 사용하였다[9]. 또한 사전학습 조건의 변화에 따른 결과의 민감도를 확인하기 위하여 MPI-Sintel 관련 체크포인트[10]를 이용한 보조 분석을 수행하였다. FlyingThings3D는 광류와 장면 흐름(Scene flow) 학습을 위해 구축된 대규모 합성 데이터셋이며[9], MPI-Sintel은 광류 알고리즘의 학습 및 평가에 널리 사용되는 벤치마크 데이터셋이다[10].

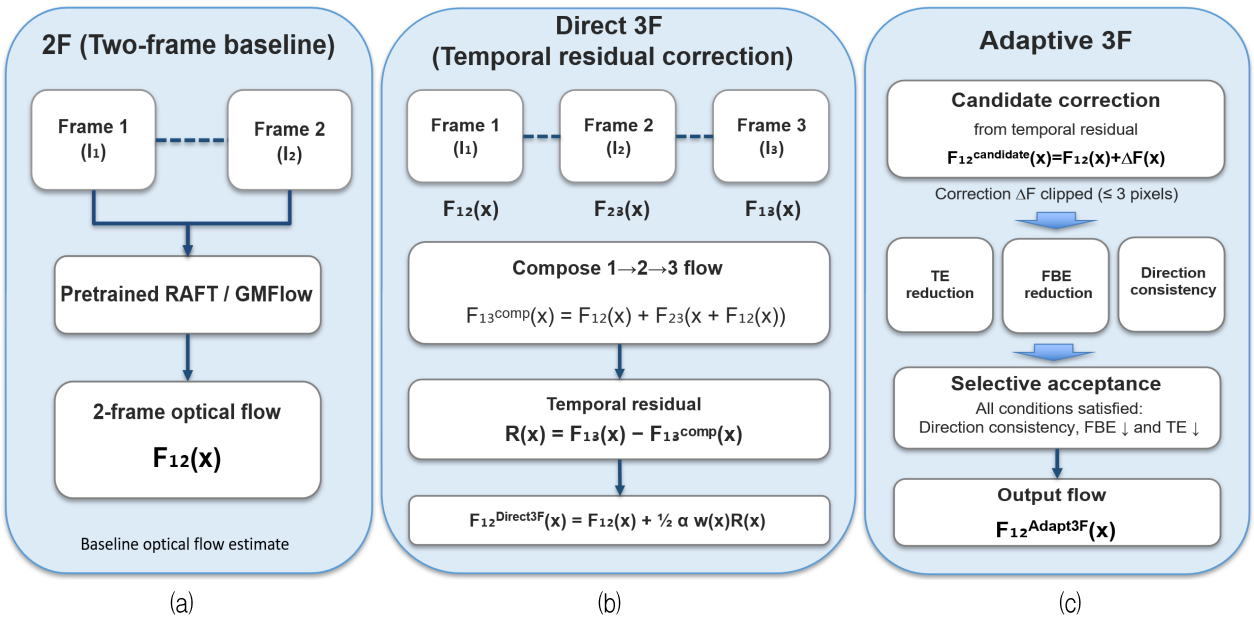
이 비교의 목적은 특정 사전학습 데이터셋의 우수성을 결정하는 것이 아니라, 본 연구에서 관찰된 시간 보정의 특성이 서로 다른 사전학습 체크포인트에서도 유지되는지를 확인하는 것으로 하였다. 주 실험에서는 RAFT의 raft-things.pth와 GMFlow의 gmflow-scale1-things.pth를 사용하였으며, 보조 분석에서는 각 모델의 Sintel 계열 체크포인트를 사용하였다.

2.3 2-프레임 기준 방법(2F)

본 연구에서 비교한 2F, Direct 3F 및 Adaptive 3F의 전체 처리 흐름을 그림 2에 나타내었다. Direct 3F는 세 프레임에서 계산한 시간 잔차를 직접 보정에 적용하는 반면, Adaptive 3F는 시간 잔차로 생성한 보정 후보 광류가 방향 일관성, FBE 감소 및 TE 감소 조건을 모두 만족할 때만 해당 후보 광류를 선택적으로 수용한다. 2F는 인접한 두 프레임 I1과 I2를 입력으로 하여 사전학습된 RAFT[6]와 GMFlow[7]에서 직접 추정한 광류 $F_{12}(x)$를 사용하는 기본 방법으로 정의하였다. 세 번째 프레임 I3의 정보나 별도의 시간적 후처리는 적용하지 않았으며, 각 사전학습 모델에서 얻은 원래의 2-프레임 광류 추정값을 해부학적 추적 성능 비교를 위한 기준으로 사용하였다.

2.4 Direct 3F 시간 보정

연속된 다중 프레임의 시간 정보를 광류 추정에 활용하는 방법들이 제안되어 왔다[11][12].



(a) (b) (c)

그림 2. 2-프레임 기준 방법(2F), Direct 3F 및 Adaptive 3F의 처리 흐름
(a) 2F: 기준 광류 추정 (b) Direct 3F: 시간 잔차의 직접 적용 (c) Adaptive 3F: 일관성이 향상될 때만 보정 수용
Fig. 2. Processing flow of the 2-frame baseline (2F), Direct 3F, and Adaptive 3F (a) 2F: baseline optical flow estimate (b) Direct 3F: temporal residual applied directly (c) Adaptive 3F: correction accepted only when consistency improves

본 연구에서는 이러한 시간적 일관성을 평가하기 위하여 세 프레임에서 $F_{12}(x)$, $F_{23}(x)$ 및 $F_{13}(x)$을 각각 계산하고, $F_{12}(x)$로 이동한 위치에서 $F_{23}(x)$을 샘플링하여 식 (3)과 같이 1→3 합성 광류를 계산하였다. 이후 직접 추정한 $F_{13}(x)$과 합성 광류 $F_{13}^{comp}(x)$의 차이를 식 (4)와 같이 시간 잔차 $R(x)$로 정의하였다.

$$F_{13}^{comp}(x) = F_{12}(x) + F_{23}(x + F_{12}(x)) \tag{3}$$

$$R(x) = F_{13}(x) - F_{13}^{comp}(x) \tag{4}$$

본 연구의 Direct 3F에서는 강건 잔차 가중치(robust residual weight), $w(x)$를 적용한 시간 잔차를 $F_{12}(x)$에 직접 더하여 식 (5)와 같이 보정하였다.

$$F_{12}^{Direct3F}(x) = F_{12}(x) + \frac{1}{2}\alpha w(x)R(x) \tag{5}$$

여기서 a는 시간 보정 강도를 조절하는 계수로 본 연구에서는 1.0으로 설정하였다. 강건 잔차 가중치 $w(x)$는 시간 잔차가 큰 영역에서 과도한 보정을 억제하기 위해 식 (6)과 같이 정의하였다.

$$w(x) = \frac{1}{1 + \left(\frac{\| R(x) \|_2}{s + \epsilon} \right)^2} \tag{6}$$

여기서 s는 유효 영역 내 시간 잔차 크기 $\| R(x) \|_2$의 90번째 백분위수이며, $\epsilon = 10^{-6}$으로 설정하였다. 즉, $w(x)$는 시간 잔차가 작은 위치에서는 1에 가까운 값을 갖고, 시간 잔차가 커질수록 감소하여 과도한 보정의 영향을 줄인다. 이 방법은 시간적 불일치를 직접 감소시키는 단순한 3-프레임 기준 방법으로 사용하였다.

2.5 Adaptive 3F 선택적 시간 보정

시간 잔차는 $F_{12}(x)$의 오차뿐 아니라 $F_{23}(x)$ 및 $F_{13}(x)$의 오차와 국소 비강체 변형을 함께 포함할 수 있다. 따라서 본 연구에서는 시간 잔차를 무조건 적용하지 않고 보정의 신뢰성과 필요성을 평가하여 선택적으로 수용하는 Adaptive 3F를 구성하였다.

먼저 시간 잔차에 강건 가중치를 적용하여 보정 후보를 생성하였다. 보정 후보 광류 $F_{12}^{candidate}(x)$는 기존 2F 광류 $F_{12}(x)$에 시간 잔차로부터 계산한 보정량 $\Delta F(x)$를 더하여 식 (7)과 같이 정의하였다. 이때 과도한 시간 보정에 따른 광류의 급격한 변화를 억제하기 위해 보정량의 크기를 최대 3 픽셀로 제한하여 식 (8)과 같이 계산하였다.

$$F_{12}^{candidate}(x) = F_{12}(x) + \Delta F(x) \tag{7}$$

$$\Delta F(x) = clip\left(\frac{1}{2}\alpha w(x)R(x),\ 3pixels \right) \tag{8}$$

여기서 $clip(\cdot)$은 보정 벡터의 방향은 유지하면서 그 크기가 3 픽셀을 초과하지 않도록 제한하는 연산을 의미한다. 이후 보정 후보의 신뢰성을 평가하기 위하여 방향 일관성, 순방향-역방향 일관성 및 시간적 일관성의 세 조건을 적용하였다. 순방향-역방향 일관성은 정방향 광류와 역방향 광류가 대응 위치에서 서로 일관되어야 한다는 조건으로, 광류의 신뢰성 및 가려짐(occlusion) 판단 등에 활용되어 왔다[13]. 이후 보정 후보와 기존 2F 광류 $F_{12}(x)$ 사이의 방향 일관성을 확인하고, 보정 후보 적용 후의 픽셀 수준 FBE와 TE가 기존 2F의 해당 오차보다 각각 감소하는 경우에만 보정을 수용하였다. 하나라도 조건을 만족하지 않는 경우에는 기존 $F_{12}(x)$를 유지하도록 식 (9)와 같이 정의하였다.

$$F_{12}^{Adapt3F}(x) = \begin{cases} F_{12}^{candidate}(x), & C_D(x) \wedge C_{FB}(x) \wedge C_T(x) \\ F_{12}(x), & otherwise \end{cases} \tag{9}$$

여기서 $C_D(x)$, $C_{FB}(x)$, $C_T(x)$는 각각 방향 일관성, FBE 감소 및 TE 감소 조건을 의미한다. $C_D(x)$는 보정 후보 광류와 기존 2F 광류 $F_{12}(x)$ 사이의 코사인 유사도(cosine similarity)가 0 이상일 때 만족하는 것으로 정의하였다. $C_{FB}(x)$는 보정 후보 광

류를 적용한 후의 FBE가 기존 2F 광류의 FBE보다 감소하는 경우를 의미한다. $C_T(x)$는 보정 후보 광류를 적용한 후의 TE가 기존 2F 광류의 TE보다 감소하는 경우를 의미한다. 세 조건을 모두 만족하는 위치에서만 보정 후보 광류를 최종 보정 광류로 채택하고, 하나라도 만족하지 않는 경우에는 기존의 2F 광류 $F_{12}(x)$를 유지하였다.

Adaptive 3F의 생성 및 보정 수용 과정에는 수동 지정 기준점을 사용하지 않았으며, 해부학적 기준점은 최종 광류의 해부학적 대응 정확도를 독립적으로 평가하기 위한 기준으로만 사용하였다. 보정 수용률은 유효 혀 영역의 전체 픽셀 수에 대한 방향 일관성, FBE 감소 및 TE 감소의 세 수용 조건을 모두 만족하여 후보 광류가 실제 수용된 픽셀 수의 비율로 산출하였다.

2.6 성능 평가 지표

WE는 추정 광류를 이용한 영상 재구성 결과와 실제 영상 사이의 광도 차이를 평가하였다. FBE는 정방향과 역방향 광류의 양방향 자기 일관성을 평가하는 지표로, 식 (10)과 같이 정의하였다[13].

$$FBE(x) = \| F_{12}(x) + F_{21}(x + F_{12}(x)) \|_2 \quad (10)$$

여기서 F_{21}은 두 번째 프레임에서 첫 번째 프레임으로의 역방향 광류이며, $F_{21}(x+F_{12}(x))$는 정방향 광류 $F_{12}(x)$에 의해 이동된 위치에서 이중선형 보간한 역방향 광류를 의미한다. FBE가 작을수록 순방향과 역방향 광류의 일관성이 높음을 의미하며, 최종 FBE는 유효 혀 영역에 포함된 픽셀별 FBE의 평균값으로 산출하였다. 다중 프레임의 시간적 정보를 이용한 광류 일관성은 기존 연구에서도 활용되어 왔으며[11][12], 본 연구에서는 TE를 식 (11)과 같이 1→2와 2→3의 합성 광류와 직접 계산한 1→3 광류의 차이로 정의하였다.

$$TE(x) = \| F_{13}(x) - F_{13}^{comp}(x) \|_2 \quad (11)$$

TE가 작을수록 세 프레임에서 추정한 광류의 시간적 일관성이 높음을 의미하며, 최종 TE는 유효 혀 영역에 포함된 픽셀별 TE의 평균값으로 산출하였다. 그러나 TE는 실제 해부학적 위치의 정확도를 직접 측정하지 않으므로, 본 연구에서는 해부학적 기준점 오차(LE)를 해부학적 추적 성능의 주요 평가 지표로 사용하고, WE는 광도 재구성 지표로, FBE와 TE는 광류 일관성 지표로 활용하였다.

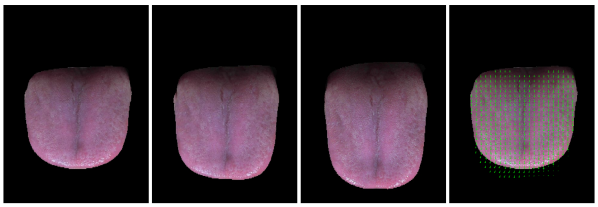
본 연구는 네 개의 3-프레임 영상 세트를 대상으로 한 연구이며, 동일 영상 세트 내의 9개 기준점은 서로 상관된 관측값이므로 이를 독립적인 표본으로 간주한 통계적 우월성 검정은 수행하지 않았다. 따라서 영상 세트별 결과와 네 영상 세트의 평균을 중심으로 해석하였다.

III. 결 과

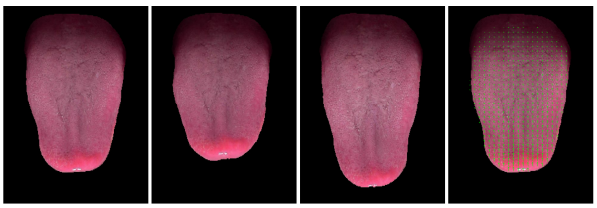
3.1 Things 체크포인트에서의 정량적 비교

본 연구에서 분석한 T1, T2, T3 및 T4의 네 개의 3-프레임 영상 세트와 각 영상 세트의 첫 번째와 두 번째 프레임 사이에서 RAFT-Things 2F로 추정한 광류 F_{12}를 그림 3에 나타내었다. LE 평가에 앞서, 9개 기준점 중 대표적인 5개 기준점에 대해 수동 기준점 지정의 관찰자 내 반복성을 평가하였다. 그 결과, 전체 기준점 지정 오차는 1.59±0.54 픽셀로 나타났다. Things 체크포인트에서 세 방법의 정량적 성능을 비교한 결과를 표 1에 나타내었다. Direct 3F는 TE를 감소시켰으나 RAFT와 GMFlow 모두에서 LE가 증가하였다. RAFT의 LE는 2F의 4.92 픽셀에서 Direct 3F의 6.24 픽셀로 증가하였고, GMFlow에서는 8.30 픽셀에서 8.85 픽셀로 증가하였다. 반면 Adaptive 3F에서는 RAFT 4.91 픽셀, GMFlow 8.24 픽셀로 나타나, Direct 3F에서 발생한 해부학적 추적 성능 저하를 억제하면서 2F와 유사한 수준을 유지하였다.

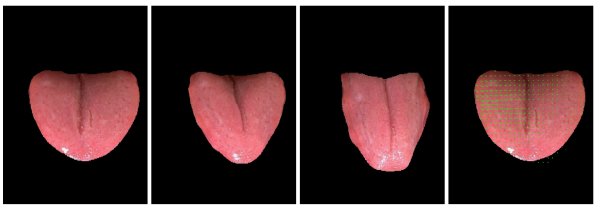
Adaptive 3F에서 세 가지 수용 조건을 모두 만족하여 보정이 실제 적용된 픽셀의 비율은 Things 체크포인트에서 RAFT 61.36%, GMFlow 46.65%였다. 이는 Adaptive 3F가 유효 혀 영역의 모든 픽셀에 일률적으로 보정을 적용하는 것이 아니라, 일관성 조건을 만족하는 픽셀에 대해서만 선택적으로 보정을 적용했음을 보여준다.



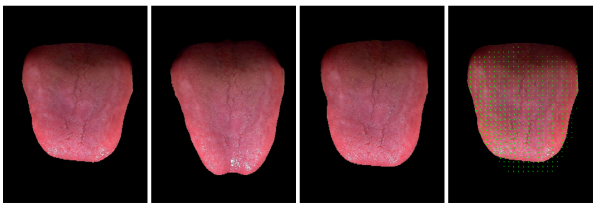
(a) T1



(b) T2



(c) T3



(d) T4

그림 3. 네 개의 3-프레임 영상 세트(I1, I2, I3)와 RAFT-Things 2F로 추정한 F_{12} 광류

Fig. 3. Four three-frame tongue image sets (I1, I2, I3) and F_{12} optical flow estimated using RAFT-Things 2F

표 1. Things 체크포인트에서의 정량적 성능 비교

Table 1. Quantitative performance comparison using the Things checkpoint

Model	Method	WE ↓	FBE ↓	TE ↓	LE ↓ (px)
RAFT	2F	18.48	5.58	7.08	4.92
	Direct 3F	19.18	4.75	4.99	6.24
	Adaptive 3F	18.45	5.37	7.03	4.91
GMFlow	2F	19.00	3.37	11.12	8.30
	Direct 3F	19.41	2.67	7.75	8.85
	Adaptive 3F	18.97	3.12	10.97	8.24

3.2 3-프레임 영상 세트별 기준점 오차

Things 체크포인트에서 각 3-프레임 영상 세트별 해부학적 기준점 오차를 표 2에 나타내었다. RAFT의 사례별 결과에서 T3는 다른 영상 세트에 비해 기준점 오차가 크게 나타났고, Direct 3F의 과도한 보정에 따른 오차 증가가 특히 뚜렷하였다. T3의 기준점 오차는 2F의 14.08 픽셀에서 Direct 3F의 19.73 픽셀로 증가하였으나, Adaptive 3F에서는 14.07 픽셀로 2F와 유사한 수준을 유지하였다. GMFlow에서도 T3의 기준점 오차가 15.18 픽셀에서 Direct 3F의 18.33 픽셀로 증가하였으며, Adaptive 3F에서는 15.00 픽셀로 나타났다. 다른 사례에서도 Adaptive 3F는 대부분 2F와 유사한 수준의 기준점 오차를 유지하였다.

표 2. Things 체크포인트에서의 3-프레임 영상 세트별 기준점 오차

Table 2. Landmark errors for individual image triplets using the Things checkpoint

Model	Case	2F	Direct 3F	Adaptive 3F
RAFT	T1	1.69	1.52	1.69
	T2	1.60	1.61	1.59
	T3	14.08	19.73	14.07
	T4	2.30	2.11	2.30
GMFlow	T1	4.73	4.72	4.68
	T2	4.30	3.80	4.30
	T3	15.18	18.33	15.00
	T4	9.00	8.54	8.99

3.3 Sintel 체크포인트를 이용한 보조 분석

사전학습 체크포인트에 따른 결과의 민감도를 확인하기 위하여 Sintel 체크포인트를 이용한 보조 분석을 수행하였으며, 그 결과를 표 3에 나타내었다. Things 체크포인트에서와 유사하게 Direct 3F는 TE를 감소시켰으나, 해부학적 기준점 오차는 2F에 비해 증가하는 경향을 보였다. 반면 Adaptive 3F는 Direct 3F에서 나타난 기준점 오차의 증가를 억제하며 2F와 유사한 수준을 유지하였다.

RAFT에서 기준점 오차는 2F의 2.31 픽셀에서 Direct 3F의 3.77 픽셀로 증가한 반면, Adaptive 3F에서는 2.30 픽셀로 나타났다. GMFlow에서도 각각 9.71, 10.52, 9.72 픽셀로 유사한 양상이 관찰되었다.

한편 체크포인트 변경에 따른 기준점 오차의 변화는 모델별로 서로 다르게 나타났다. RAFT 2F의 기준점 오차는 Things 체크포인트의 4.92 픽셀에서

Sintel 체크포인트의 2.31 픽셀로 감소한 반면, GMFlow 2F에서는 8.30 픽셀에서 9.71 픽셀로 증가하였다. 따라서 본 실험 범위에서는 특정 사전학습 체크포인트가 혀 영상에서 일관되게 우수하다고 판단하기 어려우며, 체크포인트에 따른 성능 변화가 모델에 따라 다르게 나타날 수 있음을 보여준다.

Sintel 체크포인트에서 Adaptive 3F의 보정 수용률은 RAFT 65.38%, GMFlow 63.39%로 나타났다.

표 3. Sintel 체크포인트를 이용한 보조 분석 결과
Table 3. Auxiliary analysis using the Sintel checkpoint

Model	Method	WE ↓	FBE ↓	TE ↓	LE ↓ (px)
RAFT	2F	16.12	2.20	9.62	2.31
	Direct 3F	16.93	2.54	6.99	3.77
	Adaptive 3F	16.01	2.01	9.39	2.30
GMFlow	2F	18.41	2.47	7.96	9.71
	Direct 3F	19.07	1.91	5.36	10.52
	Adaptive 3F	18.46	2.25	7.80	9.72

3.4 시간적 일관성과 해부학적 정확도

Direct 3F에서는 TE 감소와 LE 증가가 동시에 관찰되었다. RAFT-Things에서 TE는 7.08에서 4.99로 감소했지만 LE는 4.92에서 6.24 픽셀로 증가하였다. RAFT-Sintel에서도 TE는 9.62에서 6.99로 감소한 반면 LE는 2.31에서 3.77 픽셀로 증가하였다. GMFlow-Things에서도 TE는 11.12에서 7.75로 감소한 반면 LE는 8.30에서 8.85 픽셀로 증가하였으며, GMFlow-Sintel에서도 TE는 7.96에서 5.36으로 감소한 반면 LE는 9.71에서 10.52 픽셀로 증가하였다. 이는 세 프레임에서 추정된 광류의 시간적 일관성이 향상되더라도 실제 해부학적 대응 정확도가 함께 향상되는 것은 아님을 보여준다.

Ⅳ. 고 찰

본 연구의 첫 번째 주요 관찰은 시간적 일관성과 해부학적 대응이 동일한 정확도 개념이 아니라는 점이다. 기존의 다중 프레임 광류 연구에서는 연속 프레임의 시간적 정보를 이용하여 광류 추정의 일관성을 향상시키는 방법들이 제안되어 왔다[11][12]. 본 연구에서 Direct 3F는 시간 잔차를 이용하여 세 프레임 사이의 불일치를 줄였지만, RAFT와 GMFlow의 Things 및 Sintel 체크포인트를 이용한 모든 비교 조건에서 2F보다 LE가 증가하였다. 따라서 비강체 혀 움직임에서는 낮은 TE만으로 광류의 해부학적 정확도를 판단하기 어렵다.

이 현상은 혀 영상의 시각적·기하학적 특성과 관련될 수 있다. 혀 표면에서는 타액 등에 의해 정반사가 발생할 수 있으며, 이러한 반사는 혀 영상의 안정적인 특징 추출을 방해할 수 있다[8]. 혀 표면의 국소적으로 유사한 시각적 특성과 정반사는 프레임 사이의 광도 기반 대응을 어렵게 할 수 있다. 또한 혀는 비강체 구조로서 위치에 따라 서로 다른 움직임과 변형이 발생할 수 있다[1]. 이러한 특성으로 인해 F_{12}, F_{23} 및 F_{13} 각각에서 대응 오차가 발생할 수 있으며, 여러 오차가 혼합된 시간 잔차를 F_{12}에 직접 적용하면 시간적 일관성은 개선되더라도 이미 정확했던 해부학적 대응을 손상시킬 수 있다.

두 번째 주요 관찰은 Adaptive 3F가 2F보다 기준점 정확도를 크게 향상시키기보다는 Direct 3F의 해부학적 추적 성능 저하를 억제한다는 점이다. 순방향-역방향 일관성은 광류의 신뢰성 및 대응의 일관성을 평가하는 데 활용되어 왔으며[13], 본 연구에서는 이를 시간적 일관성 및 방향 일관성과 함께 보정 수용 조건으로 사용하였다. 이는 제안 방법의 목적을 '추가 프레임을 이용한 일률적인 정확도 향상'보다 '불확실한 시간 보정의 선택적 거부'로 해석하는 것이 적절함을 의미한다. 특히 2F 추정값이 이미 신뢰할 수 있는 위치에서는 보정을 적용하지 않는 것 자체가 올바른 결정일 수 있다.

세 번째로, 본 연구에서 사전학습된 RAFT와 GMFlow[6][7]를 사용한 것은 픽셀 수준의 밀집 기준 광류 구축의 현실적 어려움과 직접적으로 관련된다. 실제 혀 표면의 모든 픽셀을 연속 프레임에서 동일한 물리적 위치로 추적하여 기준 광류를 구축하는 것은 어렵고, 희소 기준점 지정만으로는 밀집 지도학습을 위한 기준 데이터를 제공할 수 없다. 이러한 이유로 광류 연구에서는 FlyingThings3D와 같은 합성 데이터셋을 이용하여 밀집 기준 광류를 확보하고 학습에 활용해 왔다[9]. 따라서 대규모 벤치마크 데이터로 사전학습된 모델의 움직임 표현을

활용하는 것은 제한된 생체의학 데이터 환경에서 적용 가능한 접근이다. 동시에 이러한 사전학습된 표현이 혀 영상에 최적화되어 있지는 않으므로 독립적인 해부학적 평가가 필요하다.

Things와 Sintel 체크포인트의 결과가 RAFT와 GMFlow에서 서로 다르게 나타난 것도 이러한 도메인 전이(Domain transfer) 문제와 관련될 가능성이 있다. 학습 기반 광류 모델은 학습 데이터와 특성이 다른 혀 영상에 적용할 경우 성능이 저하될 수 있으며, 이를 해결하기 위한 광류 도메인 적응(Domain adaptation) 연구도 이루어지고 있다[14]. Things와 Sintel은 서로 다른 영상 특성과 학습 조건을 가지므로[9][10], 혀 영상에서의 성능은 단순히 사전학습 데이터셋의 종류만으로 결정되기보다 모델 구조, 정합 방식, 학습 과정 및 영상 특성 간 상호작용에 영향을 받을 수 있다. 본 연구의 표본 수로 그 원인을 규명할 수는 없으므로, 향후 혀 영상 특화 미세조정(Fine-tuning) 또는 도메인 적응 연구가 필요하다. 구체적으로는 실제 혀 영상에 대해 광도 일관성, 순방향-역방향 일관성 및 시간적 일관성을 이용한 비지도 또는 자기지도 방식으로 사전학습 모델을 미세조정하는 방법을 고려할 수 있다. 또한 소수의 수동 해부학적 기준점을 보조적인 감독 정보로 활용하여, 혀 영상의 비강체 변형과 국소적인 시각 특성에 적응하도록 모델을 최적화하는 방법도 검토할 수 있다.

평가 관점에서도 해부학적 기준점의 역할이 중요하다. WE는 광도 기반 재구성 정확도, FBE는 양방향 자기 일관성[13], TE는 다중 프레임 시간적 일관성[11][12]을 나타내지만, 이러한 일관성 지표가 실제 해부학적 대응 정확도를 직접 보장하기는 어렵다. 특히 Direct 3F는 시간 잔차를 보정에 직접 사용하므로 TE의 감소는 방법의 구성상 부분적으로 기대되는 결과이다. 따라서 혀와 같은 생체의학적 비강체 움직임에서는 일관성 지표와 함께 독립적인 해부학적 기준을 이용한 평가가 필요하다. 또한 대표적인 5개 기준점에 대한 관찰자 내 기준점 지정 오차는 1.59±0.54 픽셀로 나타나, 수동 기준점 지정에 따른 반복 측정 변동 수준을 정량적으로 확인하였다.

본 연구에는 제한점이 있다. 첫째, 네 개의 3-프레임 영상 세트만을 이용한 연구로 결과의 일반화에는 한계가 있다. 둘째, 수동 기준점의 관찰자 내 반복성은 대표적인 5개 기준점에 대해 평가하였으나, 전체 9개 기준점에 대한 반복성과 관찰자 간 재현성은 평가하지 못하였다. 셋째, 9개 기준점은 동일한 3-프레임 영상 세트 내에서 상관된 관측값이므로 독립 표본으로 간주한 통계적 우월성 검정을 수행하지 않았다. 넷째, 설근, 중앙, 설변 및 설첨의 영역별 특성을 충분히 분석하지 못하였다. 다섯째, 픽셀 수준의 밀집 기준 광류가 없기 때문에 전체 광류장의 절대적인 해부학적 정확도를 직접 검증할 수 없었다. 향후에는 더 많은 대상자와 연속 영상을 이용한 연구, 전체 기준점 및 다수 평가자를 대상으로 한 반복성·재현성 평가, 보다 조밀한 대응점 지정 또는 합성·물리적 기준을 이용한 밀집 기준 광류 구축, 그리고 사전학습된 모델의 도메인 적응 및 혀 영상 특화 미세조정을 검토할 필요가 있다.

V. 결 론

본 연구에서는 픽셀 수준의 밀집 기준 광류 구축이 어려운 비강체 혀 영상에서 사전학습된 RAFT와 GMFlow를 이용하여 2F, Direct 3F 및 Adaptive 3F를 비교하였다. Things 체크포인트에서 Direct 3F 적용 시 RAFT의 TE는 7.08에서 4.99로 감소하였으나 LE는 4.92에서 6.24 픽셀로 증가하였고, GMFlow에서도 TE는 11.12에서 7.75로 감소한 반면 LE는 8.30에서 8.85 픽셀로 증가하였다. 반면 Adaptive 3F의 LE는 RAFT 4.91 픽셀, GMFlow 8.24 픽셀로 2F와 유사한 수준을 유지하였다. Sintel 체크포인트에서도 RAFT의 LE는 2F, Direct 3F 및 Adaptive 3F에서 각각 2.31, 3.77 및 2.30 픽셀, GMFlow에서는 각각 9.71, 10.52 및 9.72 픽셀로 유사한 경향이 확인되었다. 이러한 결과는 시간적 일관성의 향상이 실제 해부학적 대응 정확도의 향상을 의미하지 않으며, Adaptive 3F의 선택적 보정이 Direct 3F에서 발생하는 해부학적 추적 성능 저하를 억제할 수 있음을 보여준다.

또한 Things와 Sintel 체크포인트에서 RAFT와 GMFlow의 성능 변화가 서로 다르게 나타나, 사전학습된 광류 모델의 혀 영상에서의 성능이 모델과 사전학습 체크포인트에 따라 달라질 수 있음을 보여주었다. 이러한 결과는 밀집 기준 광류가 제한적

인 혀 영상에서 사전학습된 모델을 활용할 때 일관성 지표뿐만 아니라 해부학적 기준점과 같은 독립적인 평가 기준을 함께 사용하는 것이 필요하며, 추가적인 시간 정보를 신뢰도에 따라 선택적으로 적용하는 것이 중요함을 시사한다.

References

[1] K. M. Hiiemae and J. B. Palmer, "Tongue movements in feeding and speech", Critical Reviews in Oral Biology & Medicine, Vol. 14, No. 6, pp. 413-429, Nov. 2003. https://doi.org/10.1177/154411130301400604.

[2] A. Sawczuk and K. M. Mosier, "Neural control of tongue movement with respect to respiration and swallowing", Critical Reviews in Oral Biology & Medicine, Vol. 12, No. 1, pp. 18-37, Jan. 2001. https://doi.org/10.1177/10454411010120010101.

[3] D. J. Park and W. C. Lee, "Collaboration system for drone and mobile robot using TLD image tracking technique", Journal of Korean Institute of Information Technology, Vol. 16, No. 10, pp. 37-42, Oct. 2018. https://doi.org/10.14801/jkiit.2018.16.10.37.

[4] S. J. Kim, H. S. Kwon, and J. M. Kim, "A recognition of lip commands using a motion shape descriptor", Journal of Korean Institute of Information Technology, Vol. 18, No. 11, pp. 1-7, Nov. 2020. https://doi.org/10.14801/jkiit.2020.18.11.1.

[5] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion", Proc. Scandinavian Conference on Image Analysis (SCIA), Halmstad, Sweden, pp. 363-370, Jun. 2003. https://doi.org/10.1007/3-540-45103-X_50.

[6] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow", Proc. European Conference on Computer Vision (ECCV), Online, pp. 402-419, Aug. 2020. https://doi.org/10.1007/978-3-030-58536-5_24.

[7] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "GMFlow: Learning optical flow via global matching", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 8121-8130, Jun. 2022. https://doi.org/10.1109/CVPR52688.2022.00795.

[8] Z. Cui, D. Zhang, K. Wang, H. Zhang, N. Li, and W. Zuo, "Weighted nuclear norm minimization based tongue specular reflection removal", Mathematical Problems in Engineering, Vol. 2015, Art. no. 979415, Aug. 2015. https://doi.org/10.1155/2015/979415.

[9] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation", Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 4040-4048, Jun. 2016. https://doi.org/10.1109/CVPR.2016.438.

[10] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation", Proc. European Conference on Computer Vision (ECCV), Florence, Italy, pp. 611-625, Oct. 2012. https://doi.org/10.1007/978-3-642-33783-3_44.

[11] J. Janai, F. Güney, A. Ranjan, M. J. Black, and A. Geiger, "Unsupervised learning of multi-frame optical flow with occlusions", Proc. European Conference on Computer Vision (ECCV), Munich, Germany, pp. 713-731, Sep. 2018. https://doi.org/10.1007/978-3-030-01270-0_42.

[12] X. Shi, Z. Huang, W. Bian, D. Li, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li, "VideoFlow: Exploiting temporal cues for multi-frame optical flow estimation", Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, pp. 12469-12480, Oct. 2023. https://doi.org/10.1109/ICCV51070.2023.01146.

[13] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a

bidirectional census loss", Proc. AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, Vol. 32, No. 1, pp. 7251-7259, Feb. 2018. https://doi.org/10.1609/aaai.v32i1.12276.

[14] J. Yoon, S. Kim, S. Kwak, and M. Cho, "Optical flow domain adaptation via target style transfer", Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 2111-2121, Jan. 2024. https://doi.org/10.1109/WACV57701.2024.00211.

저자소개

김 근 호 (Keun Ho Kim)



1994년 2월 : 한국과학기술원 전기및전자공학과(공학사)
2001년 8월 : 한국과학기술원 전기및전자공학과(공학박사)
2001년 9월 ~ 2002년 10월 : 미네소타주립대학교 CMRR 박사후연구원
2002년 11월 ~ 2007년 5월 : 삼성종합기술원 전문연구원
2007년 5월 ~ 현재 : 한국한의학연구원 책임연구원
관심분야 : 영상처리, AI 알고리즘, 한방 의료기기, 생체신호처리