

컬러 영상 색채 강도 엔트로피를 이용한 앙상블 모델 기반의 지능형 나비 영상 인식

김태희¹ · 강승호^{2*}

Ensemble Model Based Intelligent Butterfly Image Identification Using Color Intensity Entropy

Tae-Hee Kim¹ · Seung-Ho Kang^{2*}

¹Professor, Department of Information Security, Dongshin University, Naju, 58245 Korea

^{2*}Associate Professor, Department of Information Security, Dongshin University, Naju, 58245 Korea

요 약

영상을 이용한 기계학습 기반의 나비 종 인식 기술은 나비 종의 다양성 및 개체 수, 종의 서식 분포 등을 파악하는데 관련 분야 종사자의 많은 시간과 비용 감소의 효과를 가져온다. 나비 종 분류의 정확성과 시간 효율을 높이기 위해 기계학습 모델의 입력으로 사용되는 여러 가지 특징들이 연구되었다. 그중 엔트로피 개념을 이용한 가지 길이 유사성 엔트로피나 색채 강도 엔트로피 방법이 푸리에 변환이나 웨이블릿 등 다른 특징들에 비해 높은 정확성과 적은 학습 시간을 보여주었다. 본 논문은 나비의 컬러 영상에 대한 RGB 색채 강도 엔트로피를 이용한 특징 추출 알고리즘을 제안한다. 또한 제안한 특징 추출 방법과 대표적인 앙상블 모델들을 결합한 나비 인식 시스템을 개발하고 성능을 평가한다.

ABSTRACT

The butterfly species recognition technology based on machine learning using images has the effect of reducing a lot of time and cost of those involved in the related field to understand the diversity, number, and habitat distribution of butterfly species. In order to improve the accuracy and time efficiency of butterfly species classification, various features used as the inputs of machine learning models have been studied. Among them, branch length similarity(BLS) entropy or color intensity entropy methods using the concept of entropy showed higher accuracy and shorter learning time than other features such as Fourier transform or wavelet. This paper proposes a feature extraction algorithm using RGB color intensity entropy for butterfly color images. In addition, we develop butterfly recognition systems that combines the proposed feature extraction method with representative ensemble models and evaluate their performance.

키워드 : 나비 인식, 색채 강도 엔트로피, 가지 길이 유사성 엔트로피, 앙상블 모델

Keywords : Butterfly Identification, Color Intensity Entropy, Branch Length Similarity Entropy, Ensemble Model

Received 24 May 2022, Revised 2 June 2022, Accepted 18 June 2022

* Corresponding Author Seung-Ho Kang(E-mail:drminor@dsu.ac.kr, Tel:+82-62-330-5953)

Associate Professor, Department of Information Security, Dongshin University, Naju, 58245 Korea

Open Access <http://doi.org/10.6109/jkiice.2022.26.7.972>

print ISSN: 2234-4772 online ISSN: 2288-4165

© This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright © The Korea Institute of Information and Communication Engineering.

I. 서 론

지구상에는 약 250,000종의 나비가 서식하고 있다. 유엔생물다양성과과학기구(IPBES)의 보고서[1]에 따르면 지구 온난화에 따른 기후변화나 개발에 따른 서식지 파괴 등의 이유로 유럽의 경우 나비 종의 9%가 멸종위기에 놓여있고 개체수도 30%이상 감소했다.

나비는 벌과 함께 식물 수분을 담당하는 대표적인 곤충으로 작물 생산에 직간접적 영향을 미치는 등 경제적 중요성이 어떤 곤충보다 높은 종이다. 따라서 나비 종의 개체 수를 비롯해 번식지 분포, 종 다양성 등 다양한 생물학적 지표를 파악하기 위한 기술이 매우 중요하다. 하지만 나비에 대한 전문적인 지식을 습득하고, 이를 바탕으로 나비 종을 식별할 수 있는 전문가를 양성하기 위해서는 많은 시간과 비용이 요구된다[2]. 따라서 비전문가도 사용할 수 있는 나비 종 자동 식별 도구가 개발된다면 학술 연구뿐 아니라 교육 기자재 등으로 활용될 수 있을 것이다.

최근 디지털 영상 기술과 기계학습 방법의 발전이 영상을 이용한 나비 종을 비롯한 다양한 생물 종의 자동 인식 기술을 주도하고 있다[3-6]. 영상에 대한 주성분 분석을 특징으로 사용하고 선형 분류법을 기계학습 모델로 사용한 DAISY(Digital Automatic Identification System)이 대표적이다[7]. 나비와 같은 생물 종은 종류가 다양할 뿐 아니라 형태나 색깔 등이 복잡하고 미묘한 차이가 있는 등 DAISY와 같은 기존 시스템으로는 현장에서 빠른 시간 내에 나비종을 인식하기 쉽지 않은 약점을 가지고 있다. 따라서 이러한 한계를 극복하고자 여러 특징 추출 방법 및 기계학습 방법이 제안되었다.

[8, 9]는 나비의 테두리 형태(shape)에 특정 점들을 선택하고 이들 간의 길이 분포를 엔트로피로 표현한 가지 길이 유사성 엔트로피를 특징으로 사용한 나비 인식 시스템을 제안하였다. 가지 길이 유사성 (BLS: Branch Length Similarity) 엔트로피를 사용한 시스템은 고유 열곡, 푸리에 변환, 웨이블릿 변환을 사용한 인식 시스템에 비해 짧은 학습 시간을 보장하면서도 유사한 인식률을 보여 주었다. 또한 [10]은 회색조 영상의 색채 강도 분포를 엔트로피로 정의한 색채 강도(CIE: Color Intensity) 엔트로피를 가지 길이 유사성 엔트로피와 함께 사용함으로써 정확성을 높일 수 있음을 확인하였다.

본 논문은 기존의 회색조 영상 색채 강도 엔트로피를

사용한 경우보다 인식률을 높이기 위해 나비의 컬러 영상에 대한 색채 강도 엔트로피를 정의하고 색채 강도 엔트로피 계산에 필요한 패치 분할 알고리즘을 제안한다. 제안한 특징을 입력으로 사용하는 대표적인 앙상블 모델인 랜덤 포레스트[11]와 LightGBM[12]에 기반한 나비 인식 시스템을 설계한다. 15종의 나비 영상 데이터를 사용해 정확도와 F1 점수를 기존 특징들과 비교한다.

논문의 구성은 다음과 같다. 2장에서 비교하고자 하는 두 가지 특징인 가지 길이 유사성 엔트로피와 회색조 영상 색채 강도 엔트로피에 대해 설명한다. 3장에서는 본 논문이 제안하는 컬러 영상 색채 강도 엔트로피 계산에 필요한 영상 패치 분할 알고리즘을 제안한다. 4장에서는 컬러 영상 색채 강도 엔트로피를 특징으로 사용하는 두 가지 앙상블 모델에 기반한 나비 인식 시스템을 설계한다. 5장에서는 컬러 영상 색채 강도 엔트로피를 사용했을 때의 성능을 기존 방법들과 비교하고 6장에서 결론을 맺는다.

II. 관련 연구

2.1. 가지 길이 유사성 엔트로피

나비의 형태(shape)만을 이용해 특징으로 사용하는 방법으로 BLS 엔트로피가 있다[13]. BLS엔트로피는 그림 1과 같이 단위 가지 네트워크(UBN: Unit Branch Network)의 중심 노드로부터 다른 노드까지의 길이 분포를 이용해 계산된다. 각 노드 i 에 대해 길이 비율 p_i 는 중심 노드 c 와 각 노드와의 거리 합에 대한 노드 i 와 중심 노드 c 와의 거리(l_i) 비로 구해진다 (식 1). 그리고 주어진 UBN에 대한 BLS 엔트로피 s 는 중심 노드와 노드들 간의 거리 분포의 엔트로피로 구한다.

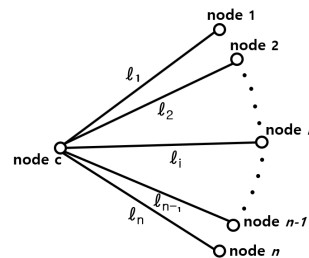


Fig. 1 Unit Branch Network

$$s = - \sum_i^n \frac{p_i \log p_i}{\log(n)}, \quad p_i = \frac{l_i}{\sum_{j=1}^n l_j} \quad (1)$$

2.2. 회색조 영상 색채 강도 엔트로피

[10]은 영상 인식에 형태 정보만을 활용하는 BLS 엔트로피에 회색조 영상에 대한 색채 강도 엔트로피를 도입함으로써 인식률을 향상할 수 있음을 보여 주었다. CI 엔트로피는 주어진 영상 객체의 색채 강도를 일정 개수로 범주화하고 객체 내 픽셀들의 범주화된 색채 강도의 분포를 엔트로피로 정의한다. 그림 2의 (a)는 회색조 영상 샘플이고 (b)는 5개로 범주화된 샘플의 채색 빈도 히스토그램이다. 각 범주에 대한 p_i 는 전체 픽셀 수에 대한 해당 범주 i 에 속하는 픽셀의 비율로 계산되며 색채 강도 엔트로피 G 는 범주화된 색채의 빈도수 분포비의 엔트로피로 정의되고 0과 1사이 값을 갖는다 (식 2). 그리고 여기서 j 는 모든 색상 강도의 모든 범주를 가리키기 위해 사용하였다.

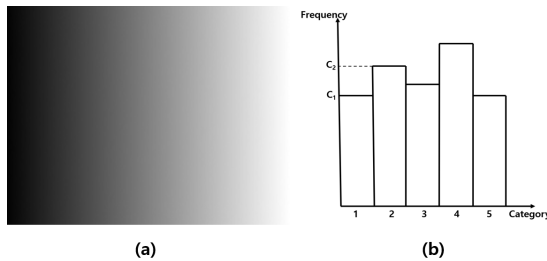


Fig. 2 (a) a patch sample of a gray image (b) the corresponding frequency graph for (a)

$$G = - \sum_i^n p_i \log p_i, \quad p_i = \frac{c_i}{\sum_{j=1}^n c_j} \quad (2)$$

III. 컬러 영상 색채 강도 엔트로피

컬러 영상을 대상으로 CI 엔트로피를 정의하는 방법과 엔트로피 계산을 위해 필요한 패치 분할 알고리즘을 나비 영상 샘플을 이용해 제시한다.

3.1. 컬러 영상 색채 강도 엔트로피

나비 영상 인식에 사용하려고 하는 컬러 영상 색채 강도 엔트로피(CG)는 기본적으로 [10]이 제시한 색채 강도 엔트로피에 기반하고 있다. 주어진 컬러 영상의 RGB 3가지 채널 각각에 색채 강도 엔트로피(G_R, G_G, G_B)를 적용하고 이들을 결합하여 사용한다. (식 3, 4, 5, 6)

$$CG = \text{concat}(G_R, G_G, G_B) \quad (3)$$

$$G_R = - \sum_i^n p_{ri} \log p_{ri}, \quad p_{ri} = \frac{c_{ri}}{\sum_{j=1}^n c_{rj}} \quad (4)$$

$$G_G = - \sum_i^n p_{gi} \log p_{gi}, \quad p_{gi} = \frac{c_{gi}}{\sum_{j=1}^n c_{gj}} \quad (5)$$

$$G_B = - \sum_i^n p_{bi} \log p_{bi}, \quad p_{bi} = \frac{c_{bi}}{\sum_{j=1}^n c_{bj}} \quad (6)$$

주어진 영상의 픽셀이 가질 수 있는 채널별 색채 강도의 범위 0부터 255까지의 값을 표1과 같이 8가지로 범주화하였다. [10]과 동일하게 8가지로 범주화한 이유는 두 가지 CI 엔트로피 이외에 범주화의 수가 성능에 영향을 미치지 않도록 하기 위해서이다.

Table. 1 Categorization of pixels according to color intensity

Category	Intensity
1	0 ~ 31
2	32 ~ 63
3	64 ~ 95
4	96 ~ 127
5	128 ~ 159
6	160 ~ 191
7	192 ~ 223
8	224 ~ 255

3.2. 영상의 패치 분할 알고리즘

나비 영상에 컬러 영상 CI 엔트로피를 적용하기 위해서는 나비 객체를 일관된 방법으로 원하는 만큼의 패치 조각으로 분할할 수 있어야 한다. 아래 나비 테두리에 등 간격으로 경계점을 사용하여 영상의 패치를 분할하는 알고리즘을 나비 영상 샘플(그림 3)과 함께 제시한다.

날개 테두리 위에서 등거리에 있는 점을 선택하기 위해서는 날개 테두리의 길이 계산이 필요하다. 아래 그림과 같이 테두리를 구성하는 연속된 두 픽셀이 위, 아래, 좌, 우로 있을 때는 거리 d 를 1로, 대각선상에 있을 때는 $\sqrt{2}$ 로 정의하여 테두리의 거리를 계산한다(그림 4).

Algorithm: 영상의 패치 분할 알고리즘

Input: 나비 영상, 패치 개수 N

Output: 픽셀과 소속 패치의 쌍 (t_i, PA_i) 집합

1. 날개 테두리 중 가장 거리가 먼 두 픽셀(t_1, t_2) 선택
2. 선분 t_1, t_2 를 등분하는 픽셀 t_3 계산
3. t_3 를 포함하고 선분 t_1, t_2 와 수직인 선분을 계산
4. 3에서 구한 선분으로부터 선분 t_1, t_2 의 20%에 해당하는 길이의 영역을 위아래로 선택
5. 4의 영역 내에서 가장 거리가 가까운 날개 테두리 위의 두 픽셀 t_4, t_5 선택
6. 선분 t_4, t_5 를 등분하는 픽셀 t_6 선택
7. t_6 를 기준으로 N 개의 등거리에 있는 픽셀(p_1, p_2, \dots, p_N)을 선택
8. 점 t_6 로부터 7에서 구해진 픽셀들을 대상으로 선분을 긋고 그 안에 놓인 나비 영역의 픽셀들에 소속 패치 부여

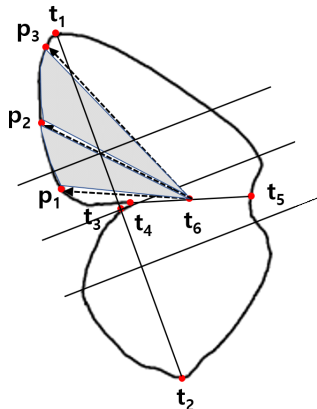


Fig. 3 patch partition sample

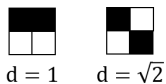


Fig. 4 Distance of consecutive two pixels in an image

패치 분할 알고리즘에 의해 사전에 정해진 N 개의 패치를 구한 후 각 패치에 소속된 픽셀들을 대상으로 R, G, B 세 채널에 대해 8가지 카테고리를 사용해 색채 강도

엔트로피를 구한다. 최종 특징으로 사용되는 데이터 형태는 이들 N 개 패치의 3개 채널에 대해 계산한 CI 엔트로피를 연결해 사용하는 것으로 특징은 $N \times 3$ 크기의 벡터가 된다. 그림 5는 N 이 5일 때의 컬러 영상 CI 엔트로피 특징 집합의 일부이다. 마지막 칼럼은 해당 특징이 소속된 나비 종을, 즉 데이터의 라벨을 나타낸다.

1.22050	1.11469	1.01125	1.34538	1.02416	1.13178	1.14635	1.11745	1.29958	1.04271	1.504878	1.469683	1.219002	1.385217	1.195693	0
1.21851	1.19274	0.95769	1.48983	1.051055	1.072492	1.27937	1.139284	1.497098	1.061563	1.350149	1.347795	1.050205	1.591724	1.33015	0
1.278309	1.420039	1.630176	1.559093	1.200309	1.250861	1.46114	1.642820	1.612534	1.204176	1.674034	1.690489	1.831982	1.575649	1.616941	0
1.189593	1.164681	1.451172	1.145127	1.02386	1.208373	1.22512	1.402931	1.15997	1.072518	1.321739	1.262405	1.431066	1.201118	1.478151	0
1.283526	1.176108	1.180165	1.357464	1.187926	1.326243	1.189308	1.224501	1.275248	1.210418	1.52773	1.407594	1.514293	1.373379	1.338387	0
0.178666	0.990238	1.948081	1.554966	0.95424	0.930266	0.90000	0.966508	1.52009	1.014534	1.046644	1.213759	1.553074	1.691902	1.134643	0
1.068162	1.101443	1.260737	1.514193	1.151482	1.015711	1.230862	1.368229	1.491408	1.114922	1.482804	1.449139	1.253621	1.465222	1.387011	0
1.070811	1.281607	1.188231	1.540731	1.106126	1.058909	1.291244	1.210043	1.520693	1.049641	1.550577	1.632142	1.606979	1.5559	1.454616	0
1.062528	0.874528	1.007205	1.483584	1.103378	0.99294	1.060719	1.49966	1.44977	0.954339	1.497896	1.552352	1.265762	1.62724	1.400884	0
1.017626	1.131686	1.295723	1.558474	1.09835	1.023378	1.141225	1.13143	1.53693	1.053905	1.488807	1.453598	1.602998	1.710808	1.347463	0
1.011822	1.539154	1.562116	1.387232	1.164512	0.962172	1.385789	1.59734	1.411632	1.08421	1.542369	1.657136	1.562738	1.732624	1.741496	0
1.045429	1.296253	1.628214	1.300912	1.06693	1.046752	1.300406	1.584004	1.348049	0.976383	1.697911	1.552487	1.452134	1.433753	0	
0.924521	0.931418	1.204205	1.449176	1.021937	0.87005	1.038412	1.322189	1.439988	0.967978	1.577038	1.524821	1.669136	1.769503	1.463823	0
0.979603	1.070588	1.163151	1.570589	1.048476	1.046564	0.99995	1.220644	1.566109	1.038447	1.718463	1.752395	1.71295	1.80526	1.701864	0
1.161601	1.511977	1.668458	1.273043	1.34709	1.190999	1.516336	1.658228	1.279158	1.308071	1.137171	1.576982	1.536278	1.651748	1.532348	0
1.783418	1.868965	1.261757	1.390609	1.899126	1.693437	1.810282	1.223848	1.373684	1.453781	1.387837	1.511096	1.075394	1.386953	1.099901	1
1.739461	1.946408	0.99861	1.399166	1.735331	1.603208	1.633873	0.947701	1.241588	1.500164	0.80735	0.923427	0.685042	0.83573	0.941065	1
1.718806	1.252141	1.380797	1.438209	1.623243	1.715026	1.59796	1.310207	1.39993	1.475106	0.896641	0.981339	0.890993	1.130922	1.093972	1
1.740638	1.820693	1.02597	1.423102	1.724881	1.669648	1.581638	1.023982	1.388569	1.429274	0.993841	0.952019	0.715597	1.118821	0.605078	1
1.811922	1.760705	1.351371	1.305036	1.759018	1.773691	1.662541	1.283872	1.227313	1.523655	1.244863	0.97312	0.851707	0.913437	0.766711	1
1.710998	1.719793	1.357179	1.492766	1.753569	1.698362	1.530036	1.252828	1.388429	1.550489	1.002145	0.807161	0.568075	0.877499	0.569626	1
1.822621	1.837962	1.329518	1.421284	1.795564	1.551067	1.214397	1.247171	1.511359	0.950296	0.761643	0.586634	0.761779	0.409274	0.77111	1
1.817961	1.85613	1.312029	1.527204	1.811842	1.693023	1.50670	1.262918	1.483005	1.58809	0.73365	0.6814	0.732396	1.02445	0.91581	1
1.800205	1.98272	1.287271	1.623042	1.893405	1.758412	1.788429	1.144017	1.470107	1.697205	0.940546	0.99488	0.854065	1.101258	0.737351	1
1.846779	1.978935	1.347889	1.564788	1.83983	1.859322	1.7399	1.250015	1.433845	1.7362	1.106501	1.017938	0.946629	1.143458	0.802812	1
1.832044	1.927935	1.261814	1.520389	1.879322	1.756817	1.667846	1.155271	1.414463	1.621774	1.127286	0.910064	0.81486	1.059863	0.907509	1
1.873384	1.551107	1.021122	1.48211	1.871633	1.688035	1.684412	1.759117	1.356568	1.584474	1.113695	1.108951	0.907759	1.045281	0.836705	1
1.737022	1.927445	1.877188	1.582111	1.738404	1.680772	1.731215	1.789961	1.50089	1.580442	1.273565	1.686882	1.228336	0.980862	1.151805	1
1.839592	1.795637	1.399759	1.529295	1.85963	1.759806	1.575745	1.924497	1.361188	1.61073	1.193849	0.989354	1.028276	0.96602	0.897202	1
1.867036	1.658427	1.210117	1.55517	1.752948	1.786299	1.362096	1.888873	1.388885	1.515787	1.127824	0.822101	0.888397	1.047047	0.804284	1
1.832455	1.77984	1.332377	1.584287	1.720176	1.623749	1.546658	1.03399	1.465791	1.479979	1.162426	1.184177	0.886097	1.21026	0.931202	1
1.832531	1.668835	1.308608	1.466637	1.720302	1.746053	1.457386	1.174023	1.374095	1.530607	1.388104	1.151368	0.970993	1.167528	0.935448	1
1.815104	1.751481	1.180028	1.2555	1.649397	1.715372	1.5477	1.954691	1.289881	1.441835	1.312608	1.048912	0.828391	1.418774	0.932818	1
1.545403	1.872941	1.835957	1.588661	1.792021	1.538014	1.70388	1.7525	1.444478	1.633274	1.2146939	1.481674	1.416087	1.130262	1.100952	1
1.701647	1.670943	1.433754	1.746273	1.850151	1.587328	1.723035	1.953582	1.589864	1.768698	1.353834	1.555521	1.546665	1.214762	1.188552	1

Fig. 5 Sample data set

IV. 앙상블 모델 기반 나비 인식 시스템

나비 영상 인식에서 컬러 영상 CI 엔트로피의 성능 효과를 확인하기 위해 앙상블 기계학습 모델을 설계한다. 앙상블 모델은 다수의 모델을 활용해 투표와 같은 방법으로 최종 의사결정을 하는 모델로서 기존 기계학습 알고리즘에 비해 우수한 성능을 보여주고 있다. 선택한 앙상블 모델은 가장 널리 알려진 랜덤 포레스트 모델 [11]과 Light GBM [12] 모델을 사용한다.

4.1. 랜덤 포레스트

랜덤 포레스트 모델은 대표적인 앙상블 모델 중 하나이며 그동안 많은 응용에 사용되어왔다. 랜덤 포레스트 모델은 결정 트리를 기본 모델로 사용하는 앙상블 모델이다. 여러 개의 결정 트리를 학습 데이터를 사용해 각각 별도로 학습시킨 후 입력 샘플에 대해 결정 트리별로 각각 분류 결정을 하게 된다. 최종 분류는 투표와 같이 다수결을 통해 결정하게 된다. 그림 6은 랜덤 포레스트의 의사 결정 과정을 보여준다.

scikit-learn이 제공하는 RandomForestClassifier를 사용하여 구현하였다. 100개의 결정트리를 사용해 랜

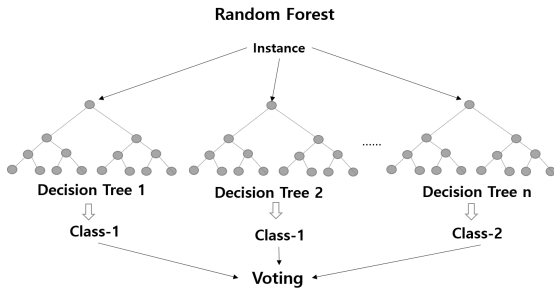


Fig. 6 Process of decision making in Random Forest model

덤 포레스트를 구성하였고, 단순 다수결 방법에 의해 최종 부류 결정을 하도록 하였다. 나머지 하이퍼 파라메타는 scikit-learn이 제공하는 기본값을 사용하였다.

4.2. LightGBM

LightGBM은 그라디언트 부스팅 결정 트리(Gradient Boosting Decision Tree: GBDT) 모델의 효율적인 구현 방법 중 하나이다. GBDT 모델은 트리 기반 앙상블 모델로 다른 기계학습 방법에 비해 높은 정확성을 보여 준다. 하지만 대량의 데이터와 많은 특징을 사용하는 경우 학습 과정에서 트리 분기에 필요한 정보 획득에 많은 시간을 요구하게 되는 문제가 있다. LightGBM은 기존의 다른 GBDT와 달리 정보 획득의 효율성을 높이기 위해 Gradient-based One-Side Sampling(GOSS) 방법과 Exclusive Feature Bundling(EFB) 방법을 도입하여 기존 GBDT의 문제점인 학습 시간의 효율성을 높일 수 있다. 트리 생성과정에서 XGBoost 등 다른 GBDT 모델이 level-wise 형식으로 트리가 생성되는 것(그림 7 (a))과 달리 LightGBM은 leaf-wise 형식으로 트리를 생성함(그림 7 (b))으로써 시간 효율성을 달성할 수 있다. 또한

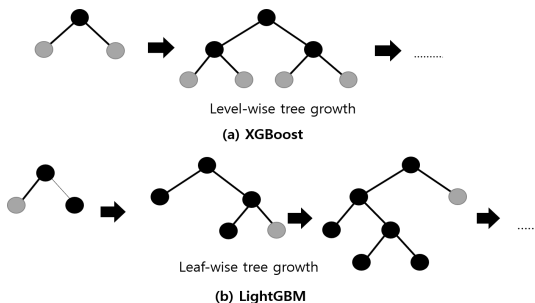


Fig. 7 The difference of constructing tree between XGBoost and LightGBM

정확성 측면에서도 XGBoost나 pGBRT와 같은 다른 기계학습 방법에 비해 낮지 않다는 것이 여러 실험 데이터를 통해 확인되었다[12].

LightGBM은 [13]이 제공한 파이썬 모듈을 사용하여 구현하였다. 랜덤 포레스트 모델과 마찬가지로 100개의 트리를 사용하여 결정하도록 하였다. 나머지 하이퍼 파라메타는 [13]에 사용된 기본값을 그대로 사용하였다.

V. 실험 및 결과 분석

15종의 나비 데이터를 이용해 모델별로 5점 교차검증 방법으로 총 20회 실행하여 평균값을 최종 성능치로 사용하였다. 분류 문제에 사용되는 모델의 성능을 평가하기 위한 성능치는 정확도를 포함해 다양한 성능치가 존재하지만 본 논문에서는 대표적으로 사용되는 정확도(accuracy)와 recall과 precision의 조화평균 값인 F1 점수를 사용하였다.

5.1. 나비 데이터

나비 데이터는 총 15가지 종으로 구성되어 있으며 이미지 총수는 423개이다. 원 영상은 760x 567 해상도를 갖는 RGB 컬러 영상이며 한쪽 날개 영상이다. 그림 8은 *Lycaena dispar* 종의 한 샘플에 대한 컬러 영상, 회색조 영상, 형태 영상을 보여 준다



Fig. 8 A sample image of *Lycaena dispar* species

5.2. 성능치

분류 모델의 성능치는 정확도와 F1 점수를 사용하였다. 우선 정확도(accuracy)는 전체 테스트 샘플에서 True Positive와 True Negative로 판별된 샘플의 비율로 계산한다(식 7). Recall은 실제 Positive인 데이터 샘플에서 True Positive로 판별한 비율로 정의되고 Precision은 모델이 Positive로 판별한 샘플 중 True Positive의 비율로

정의된다. F1 점수는 Recall과 Precision의 조화평균이다(식 8).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$F1\ score = \frac{2TP}{2TP + FP + FN} \quad (8)$$

5.3. 영상에 대한 색채 강도 엔트로피 효과 비교

우선 나비 영상을 대상으로 기존의 회색조 영상 CI 엔트로피와 컬러 영상 CI 엔트로피의 효과를 랜덤 포레스트 모델(그림 9, 10)과 LightGBM 모델(그림 11, 12)을 사용하여 비교해 보았다. 나비 영상의 패치 수를 5에서부터 30까지 변경해 가며 실험하였다. 두 모델 모두에서 컬러 영상 색채 강도 엔트로피를 사용했을 때의 정확도와 F1 점수가 10% 이상 높다는 사실을 확인할 수 있다.

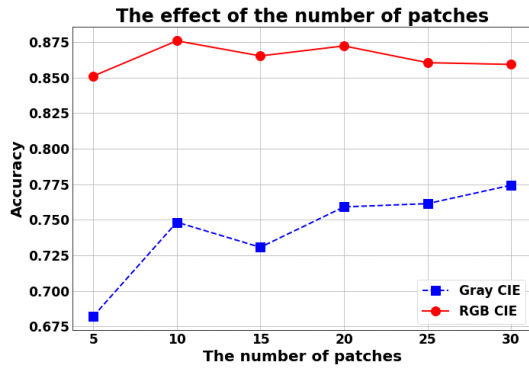


Fig. 9 The comparison of accuracy between Gray CIE and RGB CIE using random forest model

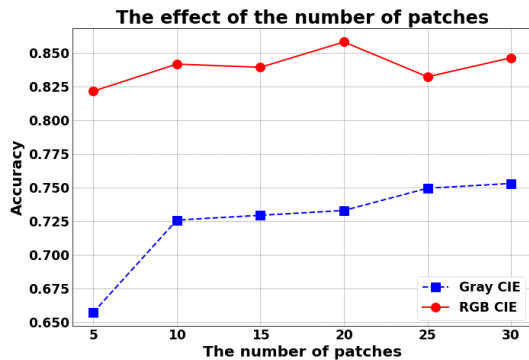


Fig. 10 The comparison of accuracy between Gray CIE and RGB CIE using LightGBM

회색조 영상 CI 엔트로피의 경우엔 패치 수가 많을수록 성능이 높아지는 것을 확인하였으나 CI 엔트로피의 경우엔 패치 수와 성능 사이에 높은 상관성은 없는 것으로 보인다. 랜덤 포레스트를 사용한 경우 패치 수가 많을 때 오히려 약간 낮은 성능을 보여 주었다. 일반적으로 모델 성능에 관해 알려진 것과 달리 색채 강도 엔트로피를 사용한 나비 영상 인식에서는 랜덤 포레스트 모델이 LightGBM 모델에 비해서 약 5% 높은 성능을 보여 주었다.

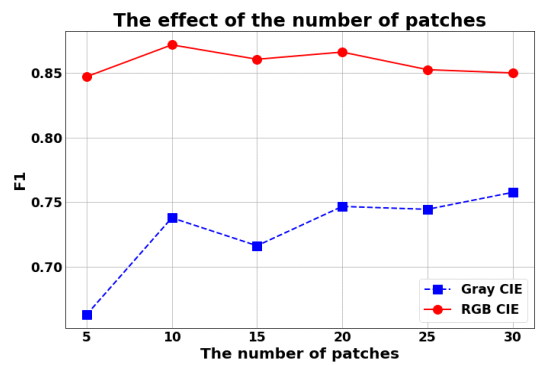


Fig. 11 The comparison of F1 score between Gray CIE and RGB CIE using random forest model

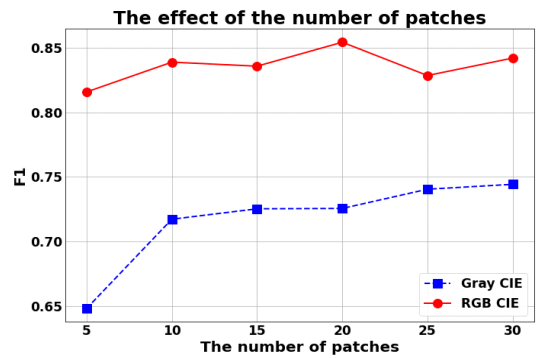


Fig. 12 The comparison of F1 score between Gray CIE and RGB CIE using LightGBM

5.4. BLS 엔트로피와 함께 사용한 경우의 효과 비교

BLS 엔트로피만 사용한 경우, BLS 엔트로피와 회색조 영상 CI 엔트로피를 함께 사용한 경우, BLS 엔트로피와 컬러 영상 CI 엔트로피를 함께 사용한 경우의 성능을 비교하였다. 회색조 영상 CI 엔트로피와 컬러 영상 CI 엔트로피의 실험과 동일하게 랜덤 포레스트 모델과

LightGBM 모델을 사용하고 정확도(그림 13, 14)와 F1 점수(그림 15, 16)를 측정하였다.

회색조 영상 CI 엔트로피와 컬러 영상 CI 엔트로피에 사용한 패치 수는 25로 고정하고 BLS 엔트로피 계산에 사용되는 점의 개수를 10에서부터 360까지 달리하여 5 겹 교차 검증을 실시하였다. 각 측정치는 20회의 실험 후 평균값을 계산한 것이다.

우선 BLS 엔트로피만 사용한 경우, 비해 CI 엔트로피를 추가하면 정확도와 F1 점수에서 두 모델 모두 10% 이상의 성능 개선이 가능함을 확인하였다. 그리고 회색조 영상 CI 엔트로피와 컬러 영상 CI 엔트로피 비교 실험에서 확인한 것처럼 BLS 엔트로피와 함께 사용한 경우에도 컬러 영상 CI 엔트로피를 사용한 경우가 더 높은 정확도와 F1 점수를 보여 주었다.

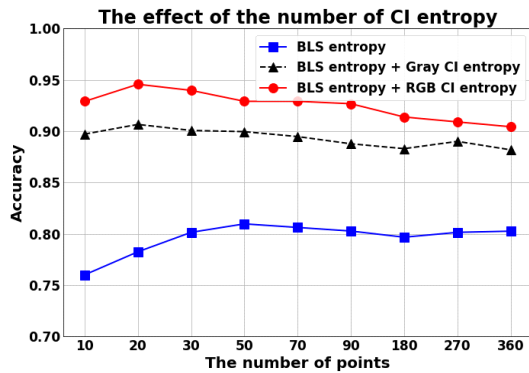


Fig. 13 The comparison of accuracy between CIEs together with BLS entropy using random forest model

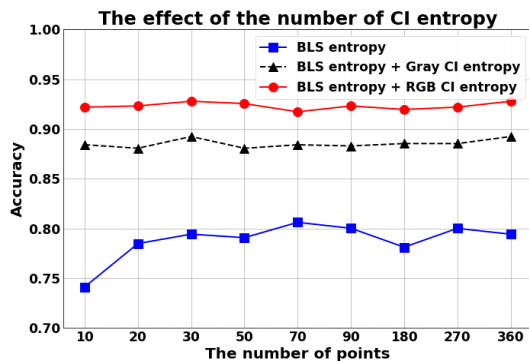


Fig. 14 The comparison of accuracy between CIEs together with BLS entropy using Light GBM

다만, BLS 엔트로피의 영향으로 두 영상 CI 엔트로피 간의 성능 차이는 단독으로 사용될 때보다 작아짐을 알 수 있다. 그리고 BLS 엔트로피 계산에 사용되는 점들의 개수는 BLS 엔트로피만을 사용한 경우가 아니라면 성능에 미치는 효과는 높지 않은 것으로 확인되었다. 따라서 속도 효율성 측면을 고려하면 적은 수의 패치와 적은 수의 점을 사용해도 충분할 것으로 생각된다.

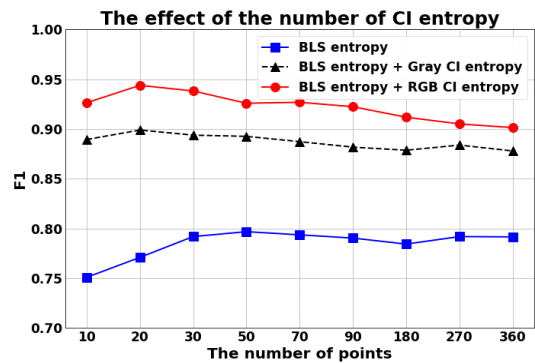


Fig. 15 The comparison of F1 score between CIEs together with BLS entropy using random forest model

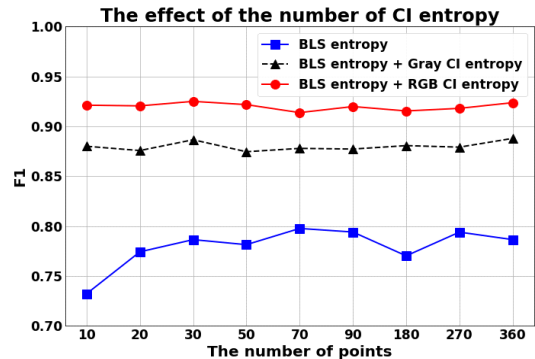


Fig. 16 The comparison of F1 score between CIEs together with BLS entropy using Light GBM

VI. 결론

본 논문은 영상을 이용한 나비 종 인식을 위해 패치 분할 알고리즘과 컬러 영상 CI 엔트로피를 제안하였다. 컬러 영상 CI 엔트로피는 Red, Green, Blue 3가지 채널에 대해 색채 강도를 8가지 카테고리로 범주화하고 이

들의 분포를 엔트로피로 수치화한 것으로 패치 분할 알고리즘에 의해 구해진 패치들을 대상으로 계산한 후 연결해 학습 및 테스트의 입력 특징으로 사용한다.

앙상블 모델인 랜덤 포레스트와 LightGBM 모델을 사용해 인식 시스템을 구현하고 5겹 교차 검증 실험을 하였다. 컬러 영상 CI 엔트로피가 회색조 영상 CI 엔트로피에 비해 정확도와 F1 점수에서 10% 이상의 우수성을 보여 주었다. BLS 엔트로피와의 특징 조합 성능 실험에서도 BLS 엔트로피를 단독으로 사용했을 때보다 10% 이상의 성능 향상을 보여 주었고 회색조 영상 CI 엔트로피와의 조합에 비해서도 약 5% 정도의 성능 우위를 확인할 수 있었다. 다만, 컬러 영상 CI 엔트로피를 사용하는 경우 특징 벡터의 크기가 커져서 대규모 데이터를 활용한 학습이나 실시간 영상 인식 등에 사용하기에는 문제가 있을 수 있다. 앞으로 컬러 영상 CI 엔트로피를 사용했을 때의 성능 보장과 속도 효율성 두 마리 토끼를 모두 잡는 방법 연구가 필요하다. 특히 UMAP이나 PCA와 같은 차원 축소를 위한 방법들을 적용하면 성능을 보장하면서 속도 효율성을 달성할 수 있을 것으로 보인다.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea under Grant NRF-2020R111A3071599.

References

- [1] IPBES, "The IPBES regional assessment report on biodiversity and ecosystem services for Europe and Central Asia," IPBES, Germany, 2018.
- [2] G. W. Hopkins and R. P. Freckleton, "Declines in the numbers of amateur and professional taxonomists: implications for conservation," *Animal Conservation*, vol. 5, no. 3, pp. 245-249, Aug. 2002.
- [3] P. J. D. Weeks and K. J. Gaston, "Image analysis, neural networks, and the taxonomic impediment to biodiversity studies," *Biodiversity&Conservation*, vol. 6, pp. 263-274, Feb. 1997.
- [4] T. Xi, J. Wang, Y. Han, C. Lin, and L. Ji, "Multiple butterfly recognition based on deep residual learning and image analysis," *Entomological Research*, vol. 52, no. 1, pp. 44-53, Jan. 2021.
- [5] D. Xin, Y. W. Chen, and J. Li, "Fine-Grained Butterfly Classification in Ecological Images Using Squeeze-And-Excitation and Spatial Attention Modules," *Applied Sciences*, vol. 10, no. 5, p. 1681, Mar. 2020.
- [6] B. A. Bakri, Z. Ahmad, and S. M. Hatim, "Butterfly Family Detection and Identification Using Convolutional Neural Network for Lepidopterology," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 2S11, pp. 635-640, Sep. 2019.
- [7] P. J. D. Weeks, M. A. O'Neill, K. J. Gaston, and I. D. Gauld, "Automating insect identification: exploring the limitations of a prototype system," *Journal of Applied Entomology*, vol. 123, pp. 1-8, Jan. 1999.
- [8] S. H. Kang, S. H. Song, and S. H. Lee, "Identification of butterfly species using a single neural network system," *Journal of Asia-Pacific Entomology*, vol. 15, no. 3, pp. 431-435, Sep. 2012.
- [9] S. H. Kang, J. H. Cho, and S. H. Lee, "Identification of butterfly based on their shapes when viewed from different angles using an artificial neural network," *Journal of Asia-Pacific Entomology*, vol. 17, no. 2, pp.143-149, Jun. 2014.
- [10] S. H. Kang and T. H. Kim, "A Performance Improvement of Automatic Butterfly Identification Method Using Color Intensity Entropy," *The Journal of the Korea Contents Association*, vol. 17, no. 5, pp. 624-632, May. 2017.
- [11] T. K. Ho, "Random Decision Forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal: QC, Canada, pp. 278-282, 1995.
- [12] G. Ke, Q. Meng, T. Finley, T. Wang, W. Ma, Q. Ye, and T. -Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach: CA, USA, pp. 3149-3157, 2017.
- [13] LightGBM [Internet]. Available: <https://lightgbm.readthedocs.io/en/latest/index.html>.



김태희(Tae-Hee Kim)

1991년 2월 동신대학교 전자계산학과 공학사
1993년 2월 전남대학교 전산통계학과 이학석사
1999년 2월 전남대학교 전산통계학과 이학박사
1998년 2월 ~ 현재 동신대학교 정보보안학과 교수
※관심분야: 정보공학, 패턴인식, 데이터베이스보안



강승호(Seung-Ho Kang)

1994년 8월 전남대학교 전산학과 이학사
2003년 8월 전남대학교 전산학과 이학석사
2009년 8월 전남대학교 전산학과 이학박사
2013년 9월 ~ 현재 동신대학교 정보보안학과 부교수
※관심분야: 알고리즘, 기계학습, 정보보안