JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Logical Consistency and Greater Descriptive Power for Facial Hair Attribute Learning

Haiyu Wu, Grace Bezold, Aman Bhatta, Kevin W. Bowyer

Department of Computer Science & Engineering, University of Notre Dame

WED-AM-035

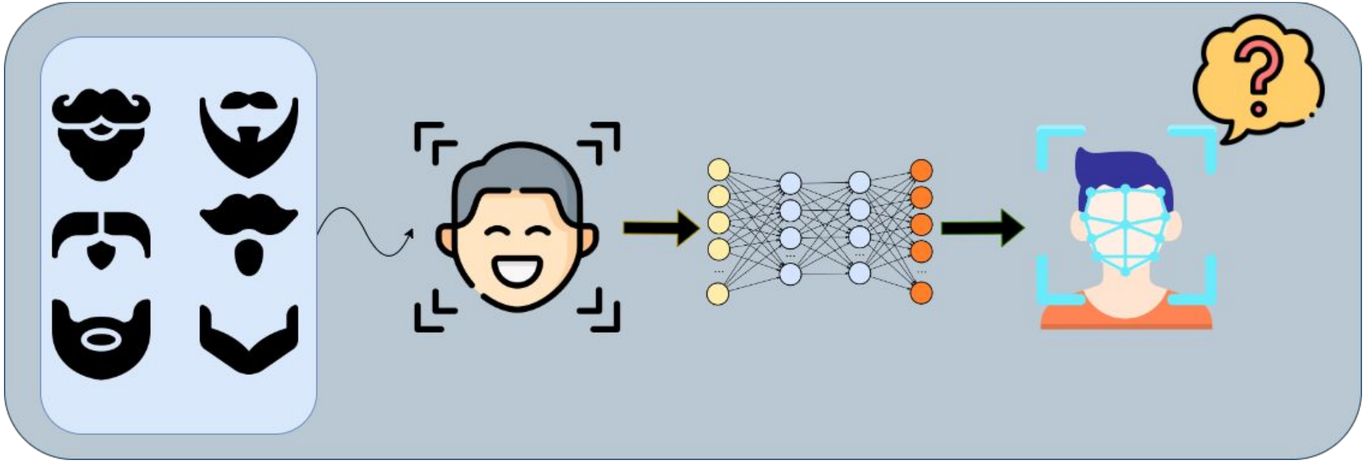
COMPUTER VISION @ND

Quick Preview

Effect of facial hair in face recognition

Facial hair attributes in existing datasets:

- **5 o'clock shadow**
- **Mustache**
- **Sideburns**
- **No Beard**
- **Goatee**



FH37K (37,000+ images with facial hair annotations)

Goal

Richer information on facial hair - area, length, connectedness

Annotation options

Beard area: clean shaven, chin area, side to side, info_not_vis

Beard length: 5 o'clock shadow, short, medium, long, info_not_vis

Mustache: none, isolated, connected to beard, info not vis

Sideburns: none, sideburns present, connected to beard, info not vis

Bald: false, top only, sides only, top and sides, info not vis

Contribution: Dataset with more descriptive facial hair annotations.

Classification performance

Performance of the models trained with FH37K dataset

- Traditional methods do not consider the logical relationships of attributes
- Methods that handle the data imbalance might give a high accuracy illusion on positive side
- Proposed LCPloss and label compensation strategy has the best performance.

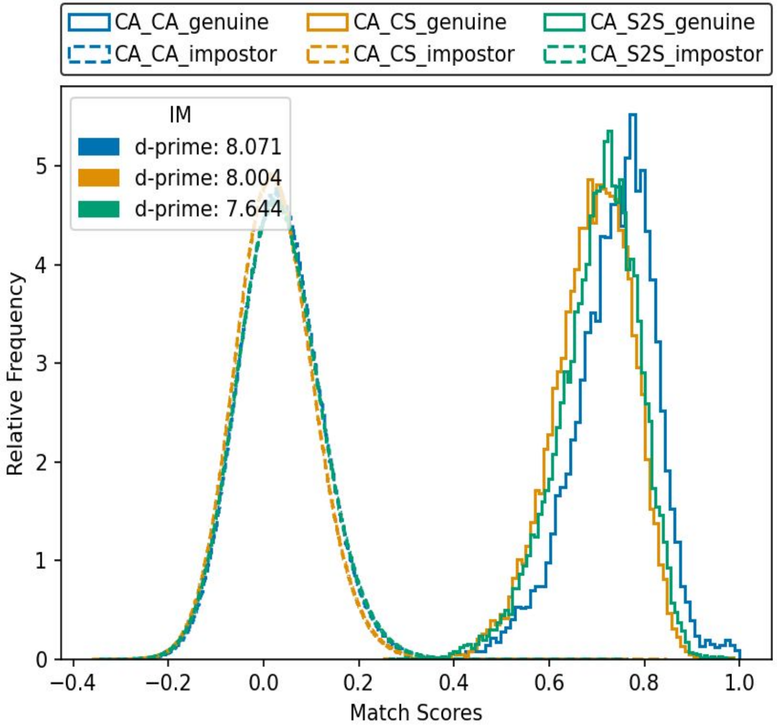
model training	ACC_{avg}	ACC^n_{avg}	ACC^p_{avg}
Not considering logical consistency ...			
BCE	88.82	93.72	54.97
BCE*	90.22	94.72	63.73
BCE-MOON*	88.96	90.67	**81.75**
BF*	89.84	95.43	58.41
Considering logical consistency ...			
BCE	45.10	46.02	32.62
BCE*	53.29	54.59	42.40
BCE-MOON*	46.46	47.54	32.95
BF*	39.96	40.95	31.45

model training	ACC_{avg}	ACC^n_{avg}	ACC^p_{avg}
Label compensation on test ...			
BCE + LC	87.47	90.08	61.55
BCE + LC*	88.83	91.49	68.78
BCE-MOON + LC*	49.39	50.55	34.62
BF + LC*	88.10	90.91	66.05
BCE + LCP + LC	87.82	90.37	59.05
BCE + LCP + LC*	89.46	92.02	66.71
Label compensation on train and test ...			
BCE + LCP + LC	88.30	91.10	62.44
BCE + LCP + LC*	**89.89**	**92.65**	**70.23**

Contribution: Approach to handle logical consistency across annotations.

Effect of beard area in face recognition

Image pairs with same beard area **increases** similarity value on **genuine side and impostor side**

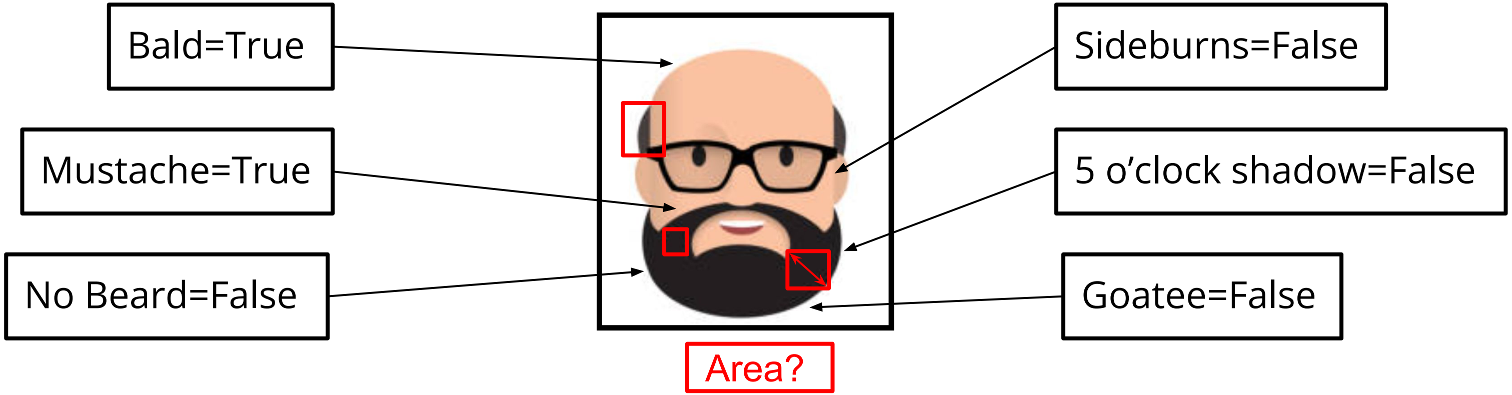


Same beard area has **different** effect in face recognition accuracy across **different** demographics.

Contribution: Facial hair effects on recognition accuracy across demographics.

FH37K dataset

Facial hair description



More descriptive attributes are needed!

Limitations of existing datasets

	# of images	# of ids	# of facial hair attributes	Area	Length	CNDN	E_c
Berkeley Human Attributes [10]⋆	8,053	-	0	0	0	0	✗
Attributes 25K [55]	24,963	24,963	0	0	0	0	✗
FaceTracer [29]⋆	15,000	15,000	1 (Mustache)	0	0	0	✗
Ego-Humans [48]	2,714	-	1 ([illegible]	[illegible]	[illegible]	0	✗
CelebA [36]⋆	202,599	[illegible]	[illegible]	[illegible]	[illegible]	0	✗
LFWA [36]⋆	13,233	[illegible]	[illegible]	[illegible]	[illegible]	0	✗
PubFig [32]⋆	58,797	[illegible]	[illegible] lock, Goatee, ...)	1	1	0	✗
LFW [26]⋆	13,233	5,749	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
UMD-AED [22]	2,800	-	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
YouTube Faces Dataset (with attribute labels [23])	3,425	1,595	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
CelebV-HQ [57]⋆	35,666 video clips	15,653	5 (5 o'Clock, Goatee, ...)	1	1	0	✗
MAAD-Face [47]⋆	3.3M	9,131	5 (5 o'Clock, Goatee, ...)	1	1	0	✓
FH37K (this paper)	37,565	5,216	**17 (Chin area, Short...)**	**4**	**4**	**4**	✓

Poor facial hair descriptions

Lac
des

Lack of evaluation on ground truth labels

Table 1. Comparison of facial hair descriptions in face attribute datasets. CNDN and E_c stand for connectedness and estimating the consistency rate of the annotations. Datasets with ⋆ are available online. FH37K has richer annotations that can cover the area, length, and connectedness of the facial hair.

FH37K

Goal

Richer information on facial hair - area, length, connectedness

Annotation options

Beard area: clean shaven, chin area, side to side, info_not_vis

Beard length: 5 o'clock shadow, short, medium, long, info_not_vis

Mustache: none, isolated, connected to beard, info not vis

Sideburns: none, sideburns present, connected to beard, info not vis

Bald: false, top only, sides only, top and sides, info not vis

000020.jpg

Data

The images we used are the subset of the CelebA, which are originally marked as No Beard = False. Also, the subset of the WebFace260M dataset (picking images for minority classes).

Documentation is available at: [Definition of Facial Hair Annotations]

Facial hair description

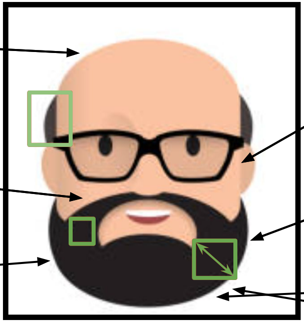
Bald: Top only=True

Bald=True

Mustache: Connected to beard=True

Mustache=True

No Beard=False



Area✅

Sideburns: None

Sideburns=False

5 o'clock shadow=False

Long=True

Goatee=False

Beard Area: Side to side=True

We have more powerful descriptions!

We need to enrich the attributes!

Logical consistency of predictions

Categories of logical relationships

- **Mutually exclusive**: The relationship among positive predictions must be **logical**, otherwise the predictions are **impossible**.

- **Dependency**: If attribute A is true, the attribute B **must be true**, otherwise the predictions are **impossible**.

- **Collectively exhaustive**: One of a group of attributes **must be true**, otherwise the predictions are **incomplete**.

Failed prediction detection

Algorithm 1 Failed prediction detection

Attribute groups

Beard areas: Clean Shaven, Chin Area, Side to Side, Info not Vis
Beard lengths: 5 O'clock Shadow, Median, Long, Info not Vis
Mustache: None, Isolated, Connected-to-beard, Info not Vis
Sideburns: None, Present, Connected-to-beard, Info not Vis
Bald: False, Top only, Sides only, Top and Sides, Info not Vis

Fail conditions

Mutually exclusive:

1. More than one positive predictions in Beard areas (except Info not Vis), Beard lengths (except Info not Vis) Mustache, Sideburns, Bald group
2. Clean Shaven + any of the Beard lengths/Mustache Connected-to-beard/Sideburns Connected-to-beard
3. Chin area + Sideburns Connected-to-beard
4. Bald (Top and Sides or Sides only) + having sideburns (Sideburns Present, Sideburns Connected-to-beard)

Dependency:

1. Having beard (Chin Area, Side to Side) + one of the beard lengths must be true
2. Mustache is connected to beard + !(Chin Area, Side to Side)
3. Sideburns is connected to beard + !Side to Side

Collectively exhaustive

No positive prediction in Beard area/Beard lengths/Mustache/Sideburns/Bald

Impossible: prediction fits any of the conditions in *Mutually exclusive* and *Dependency*
Incomplete: prediction fits any of the conditions in *Collectively exhaustive*

Logical consistency on predictions

model training	ACC_{avg}	ACC^{n}_{avg}	ACC^{p}_{avg}
Not considering logical consistency ...			
BCE	88.82	93.72	54.97
BCE*	90.22	94.72	63.73
BCE-MOON*	88.96	90.67	**81.75**
BF*	89.84	95.43	58.41
Considering logical consistency ...			
BCE	45.10	46.02	32.62
BCE*	53.29	54.59	42.40
BCE-MOON*	46.46	47.54	32.95
BF*	39.96	40.95	31.45

Backbone: ResNet50

Loss functions:
Baseline: BCE
Handling imbalance data: BF, BCE-MOON

*: transfer learning

*Every logically inconsistent prediction is considered as **incorrect***

After considering logical consistency on predictions, the accuracy drops **significantly!**
(43.26% decrease on average)

Real-world evaluation

Test set **(600K images)**:
The images in the first 30,000 ID folders of WebFace260M

model training	N_{inp}	N_{imp}	R_{failed}
BCE	333,773	1,054	55.05
BCE*	242,279	6,034	40.83
BCE-MOON*	31,656	315,756	57.12
BF*	340,898	1,314	56.27

On average, 52.32% of the predictions are failed

The proposed methods

LCPLoss - handles impossible cases

Step1: Group attributes

Mutually exclusive:

$$A_{ex} = \{attr_1, attr_2, ..., attr_N\}$$

$$L_{ex} = \{l_1, l_2, ..., l_N\}$$

Dependency:

$$A_d = \{attr_1, attr_2, ..., attr_N\}$$

$$L_d = \{l_1, l_2, ..., l_N\}$$

Step2: Conditional Probability on predictions

$$\mathcal{P}_d = \mathcal{P}(L_d|A_d) \qquad \mathcal{P}_{ex} = \mathcal{P}(A_{ex} \cap L_{ex})$$

Since $\mathcal{P}_{ex} = \mathcal{P}(L_{ex}|A_{ex})P(A_{ex})$, we can formulate the calculation of $\mathcal{P}_{ex}$ and $\mathcal{P}_d$ as:

$$\mathcal{P} = \frac{1}{N}\sum_{i=0}^{N}\mathcal{P}(\sum l_i > 0|attr_i == 1) \tag{4}$$

l_i and $attr_i$ are **binary predictions** after thresholding.

Step3: Optimization

Force $\mathcal{P}_{ex}$ to 0, $\mathcal{P}_d$ to 1

$$\mathcal{L}_{LCP} = ||\alpha\mathcal{P}_{ex} + \beta(1 - \mathcal{P}_d)||^2$$

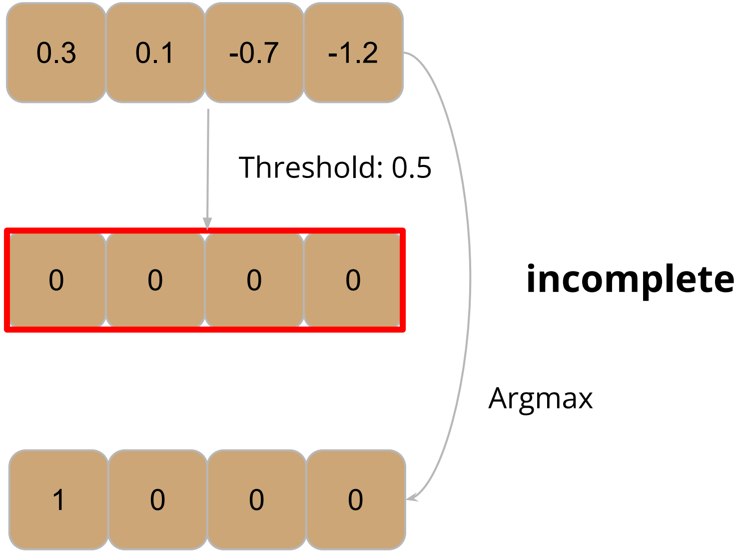
Step4: Combine with BCE

$$\mathcal{L}_{total} = (1 - \lambda)\mathcal{L}_{BCE} + \lambda\mathcal{L}_{LCP}$$

Label compensation - handle incomplete cases

Pick the **maximum** confidence value as the **positive** prediction in a **group** of attributes

Beard area: CS, CA, S2S, Info not Vis



Accuracy

model training	ACC_{avg}	ACC^{n}_{avg}	ACC^{p}_{avg}
Not considering logical consistency ...			
BCE	88.82	93.72	54.97
BCE*	90.22	94.72	63.73
BCE-MOON*	88.96	90.67	**81.75**
BF*	89.84	95.43	58.41
BCE + LCP	88.90	95.55	46.13
BCE + LCP*	90.63	95.87	58.15
BCE + LCP + LC	89.11	95.06	52.17
BCE + LCP + LC*	**90.90**	**95.98**	63.30

model training	ACC_{avg}	ACC^{n}_{avg}	ACC^{p}_{avg}
Considering logical consistency ...			
BCE	45.10	46.02	32.62
BCE*	53.29	54.59	42.40
BCE-MOON*	46.46	47.54	32.95
BF*	39.96	40.95	31.45
BCE + LCP	27.66	28.19	18.80
BCE + LCP*	42.86	43.70	33.67
Label compensation on test ...			
BCE + LC	87.47	90.08	61.55
BCE + LC*	88.83	91.49	68.78
BCE-MOON + LC*	49.39	50.55	34.62
BF + LC*	88.10	90.91	66.05
BCE + LCP + LC	87.82	90.37	59.05
BCE + LCP + LC*	89.46	92.02	66.71
Label compensation on train and test ...			
BCE + LCP + LC	88.30	91.10	62.44
BCE + LCP + LC*	**89.89**	**92.65**	**70.23**

Conclusions:

1. Label compensation can improve the accuracy
2. Labeling images in the logically consistent way can guide the model learning the logically consistent pattern on-the-fly
3. The classification method that can handle the imbalance data can give a **high-accuracy illusion**
4. **The proposed method has the outstanding performance**

Logical consistency in real-world evaluation

model training	N_{inp}	N_{imp}	R_{failed}
BCE	331,870	1,038	55.13
BCE*	240,761	6,001	40.86
BCE-MOON*	31,512	313,044	57.05
BF*	339,136	1,295	56.37
BCE + LCP	470,806	117	77.98
BCE + LCP*	307,576	300	50.98
Label compensation on test ...			
BCE + LC	0	10,215	1.69
BCE + LC*	0	11,134	1.84
BCE-MOON + LC*	0	330,115	54.66
BF + LC*	0	14,007	2.32
BCE + LCP + LC	0	14,097	2.33
BCE + LCP + LC*	0	6,083	1.01
Label compensation on train and test ...			
BCE + LCP + LC	0	7,693	1.27
BCE + LCP + LC*	0	5,595	**0.93**

Conclusions:

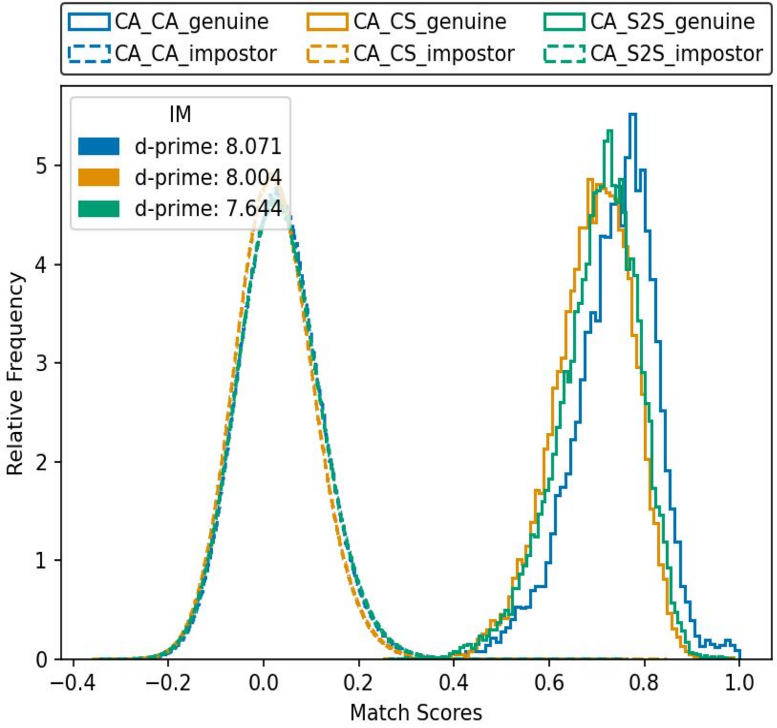
1. The Label compensation method can **eliminate** the incomplete cases
2. Labeling images in the logically consistent way can guide the model learning the logically consistent pattern on-the-fly
3. **The proposed method has the lowest fail rate**

Effect of beard area on face recognition accuracy

Chin area focused

Dataset: **BA-test**
Face matcher: MagFace

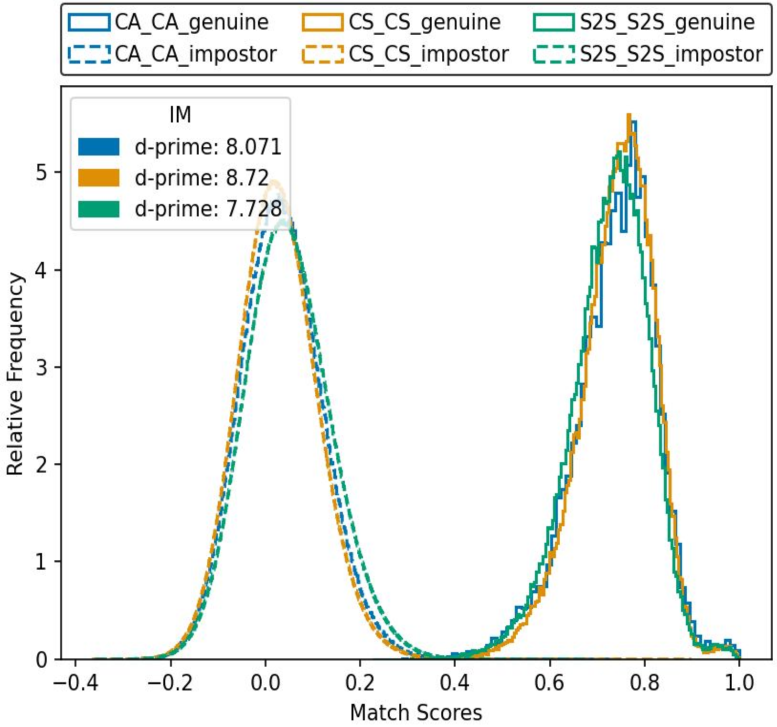
Image pairs with same beard area **increases** similarity value on both **genuine side and impostor side**



Same beard area focused

Similarity value high to low

Impostor side:
WM: (S2S,S2S) > (CA,CA) > (CS,CS)
AM: (CA,CA) > (S2S,S2S) > (CS,CS)
BM: (S2S,S2S) > (CA,CA) > (CS,CS)
IM: (S2S, S2S) > (CS,CS) > (CA,CA)



MagFace

Similarity value high to low

Genuine side:
WM: No obvious difference
AM: (S2S,S2S) > (CA,CA) > (CS,CS)
BM: (CS,CS) > (CA,CA) > (S2S,S2S)
IM: (CS,CS) == (CA,CA) > (S2S,S2S)

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Logical Consistency and Greater Descriptive Power for Facial Hair Attribute Learning

Haiyu Wu, Grace Bezold, Aman Bhatta, Kevin W. Bowyer

Department of Computer Science & Engineering, University of Notre Dame

WED-AM-035

COMPUTER VISION @ND