



DL Seminar

DeepSORT

Simple Online and Realtime Tracking with
a Deep Association Metric



한양대학교
HANYANG UNIVERSITY

인공지능 연구실
김지성

Introduction

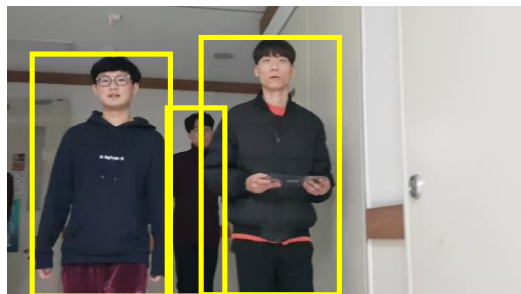
Multi Object Tracking

Frame 1

Frame 2

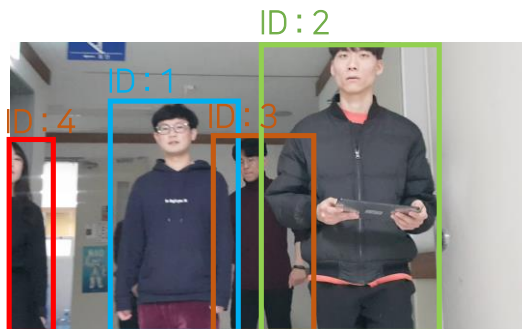
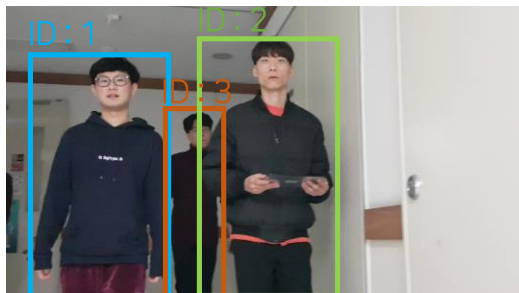
Frame N

탐지 결과



...

추적 결과

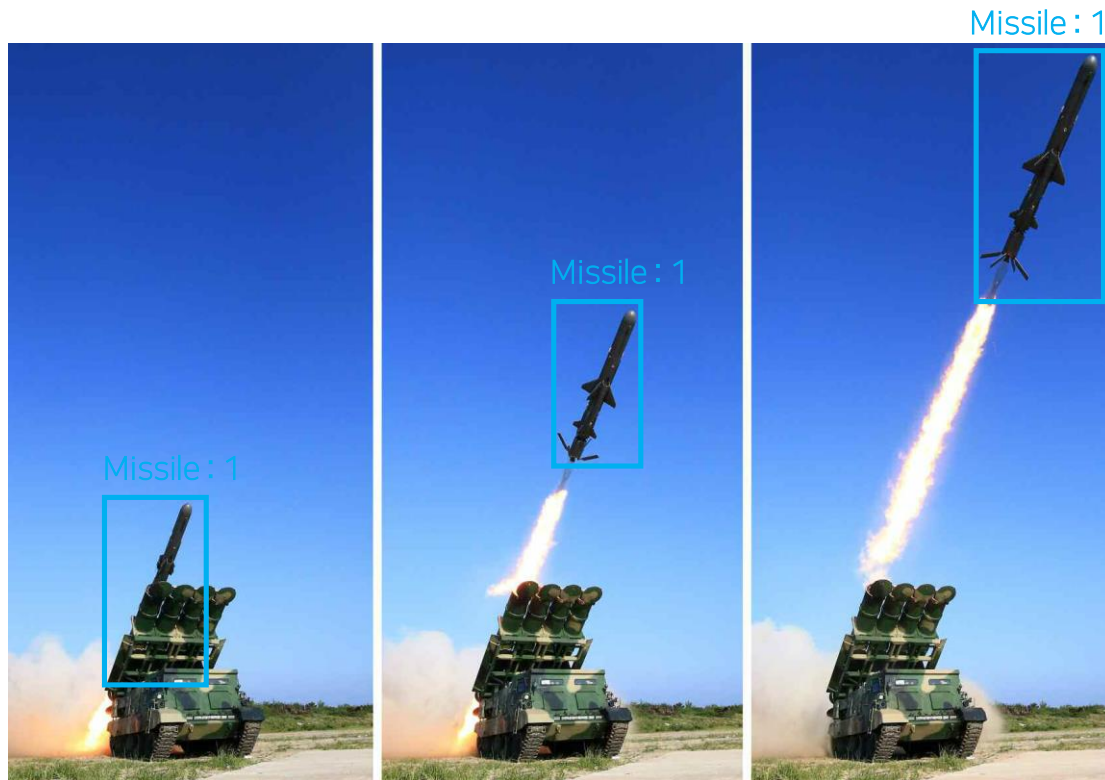


...

처음 탐지한 객체에 ID를 부여하고, 이후 프레임에서도 같은 ID로 인식하는 문제
연속적인 영상(시계열 데이터)에서 시간에 따른 객체의 공간적, 시각적 변화를 얻을 수 있다.

Introduction

Multi Object Tracking

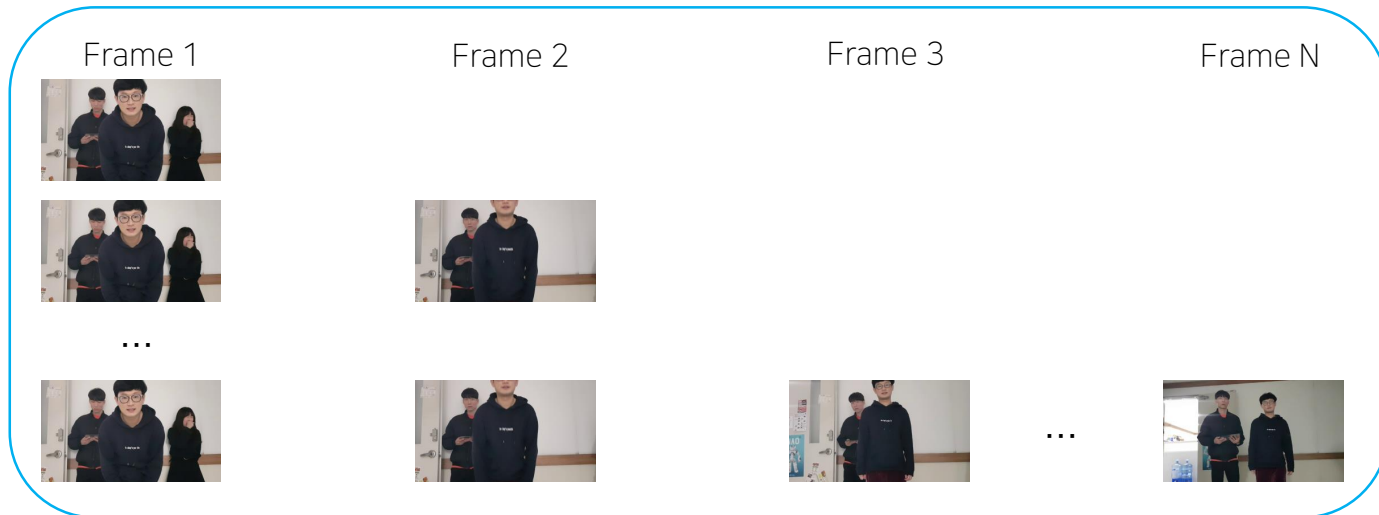


→ 다음 위치는 어디?

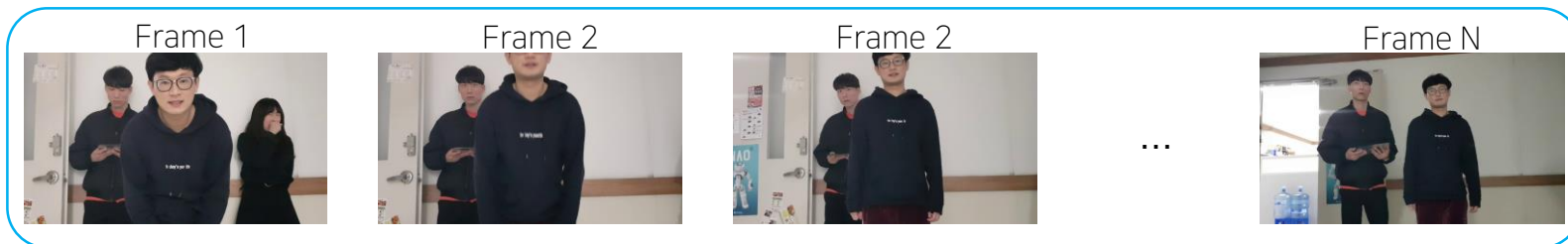
처음 탐지한 객체에 ID를 부여하고, 이후 프레임에서도 같은 ID로 인식하는 문제
연속적인 영상(시계열 데이터)에서 시간에 따른 객체의 공간적, 시각적 변화를 얻을 수 있다.

Introduction

Online vs Batch



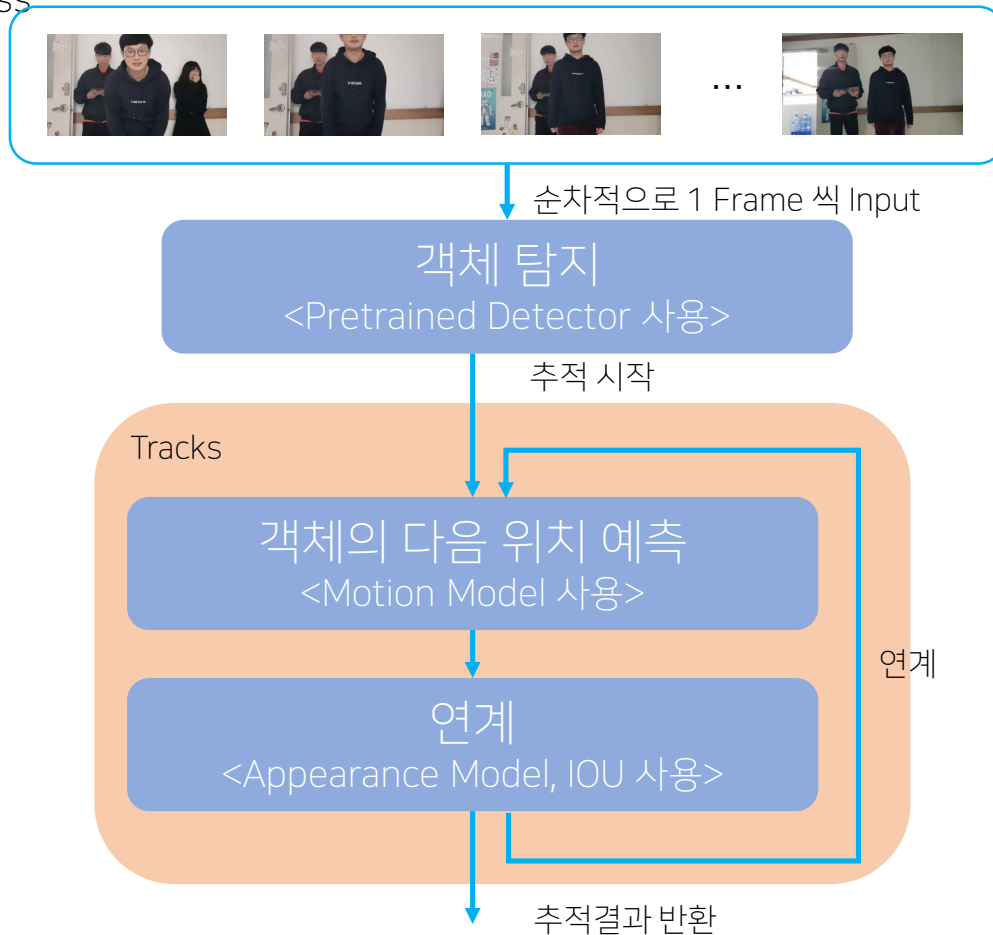
Online Tracking : 현재로부터 과거 프레임 관측 -> 실시간 추적 가능, 상대적으로 낮은 정확도



Batch Tracking : 전체 프레임 관측 -> 실시간 추적 불가능, 상대적으로 높은 정확도

Introduction

General Tracking Process



Introduction

Detect – Pretrained detector

	backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
<i>Two-stage methods</i>							
Faster R-CNN+++ [3]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [6]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [4]	Inception-ResNet-v2 [19]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [18]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	52.1
<i>One-stage methods</i>							
YOLOv2 [13]	DarkNet-19 [13]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [9, 2]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [2]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [7]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [7]	ResNeXt-101-FPN	40.8	61.1	44.1	24.1	44.2	51.2
YOLOv3 608 × 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

COCO for YOLOv3

Introduction

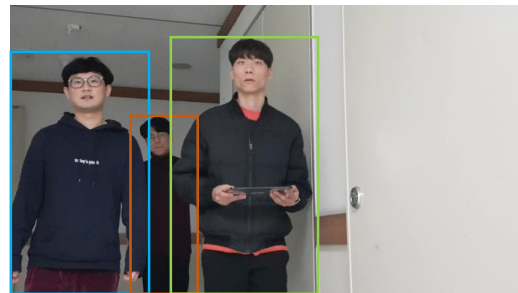
Predict - Motion model

Frame 1

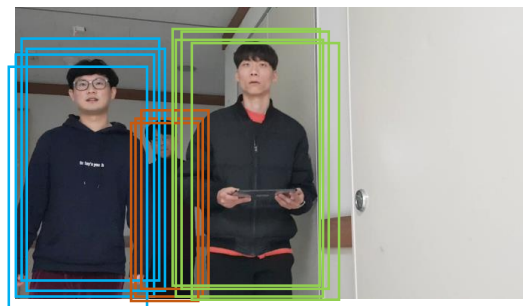
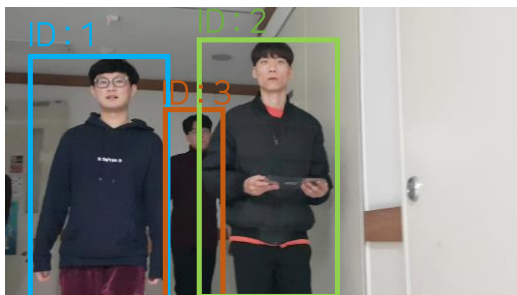


칼만필터 예측

Frame 2



파티클필터 예측



Introduction

Association

Frame 1



Frame 2



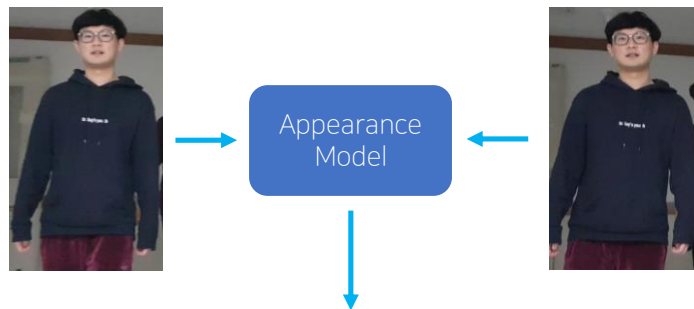
- 이전 위치
- 예측한 위치

IOU를 이용한 동일객체 연계 : Detector 성능에 매우 의존적



겹침 정도(IOU)

Appearance Model을 이용한 동일객체 연계 : Detect Noise 보완



객체간 유사도

Method

Body Embedding



A인물의 사진1



Body
Embedding



```
array([-0.07440512, 0.13833548, 0.01550988, -0.04143689, -0.11708137,  
0.017141652, -0.09219746, -0.04101015, 0.12901564, -0.03510666,  
0.26017255, -0.01175268, -0.23214489, -0.10993981, -0.06433399,  
0.14533879, -0.18374404, -0.07009007, 0.01960303, 0.03927957,  
0.17917837, 0.10840175, 0.06728961, 0.01388871, -0.0921066,  
-0.32157087, -0.06875613, -0.11010363, 0.02105946, -0.09806888,  
0.095232023, -0.01714757, -0.16107245, -0.0485564, 0.05062422,  
0.04144309, -0.03346116, -0.03011878, 0.15263124, 0.01069992,  
-0.23673587, 0.05740198, 0.03090802, 0.23846291, 0.06076686,  
0.01160336, 0.00995683, -0.15661725, 0.09741502, -0.1173826,  
0.08130169, 0.15210505, 0.14222656, 0.02321066, 0.00600557,  
-0.07693202, -0.02959017, 0.15295519, -0.13042481, 0.03233673,  
0.1088061, -0.05199346, -0.01501178, -0.08011279, 0.17588341,  
0.02202863, -0.12124902, -0.26303217, 0.06608438, -0.128976,  
0.14779539, 0.1516934, -0.16154116, -0.1716671, -0.25228465,  
0.01865603, 0.36660314, 0.04856375, -0.18907131, 0.05604936,  
-0.05154709, -0.04398936, 0.08519959, 0.14640823, 0.00993889,  
0.02739849, -0.11102622, 0.01848426, 0.23328975, -0.1351117,  
-0.04641433, 0.22538319, -0.00492372, 0.10828383, 0.02823116,  
0.02502055, -0.02946196, 0.0702677, -0.09494975, -0.0334047,  
0.00996118, -0.08729131, -0.04657208, 0.09973253, -0.14260028,  
0.10422572, -0.02379006, 0.05201333, 0.02785442, -0.09933321,  
-0.09983464, -0.02418252, 0.13822252, -0.24259827, 0.2356335,  
0.12104575, 0.14091188, 0.07011457, 0.10868125, 0.04752614,  
0.02115094, -0.04581762, -0.23486597, -0.0105925, 0.11252803,  
-0.05433004, 0.10194337, -0.02596316])
```



Euclidean distance

0.78



A인물의 사진2



Body
Embedding



```
array([-0.08902847, 0.14762324, 0.05718813, -0.09012994, -0.09676401,  
0.02029476, -0.08940992, -0.038939, 0.01385386, -0.03121,  
0.2729257, -0.02359607, -0.23816103, -0.09732635, -0.03333118,  
0.1333721, -0.19538365, -0.0766579, 0.02075577, 0.03931012,  
0.15297598, 0.08741103, 0.057871, 0.0144625, -0.0895172,  
-0.33090472, -0.06792867, -0.10570621, 0.01662679, -0.11775655,  
-0.10916033, -0.04665562, -0.1787585, -0.05185516, 0.04875493,  
0.0426253, -0.0591871, -0.04656718, 0.16380528, 0.00791238,  
-0.22012679, 0.04296945, 0.04917528, 0.23999002, 0.21356051,  
0.00689229, 0.0836444, -0.11314141, 0.09843643, -0.162096,  
0.11051578, 0.15530622, 0.12857951, 0.05687075, 0.01556353,  
-0.05330104, -0.04205765, 0.14811578, -0.10172325, 0.06627692,  
0.11120091, 0.05348025, -0.0242785, -0.0697246, 0.14077779,  
0.04438576, -0.13614309, -0.27176958, 0.04076207, -0.10192671,  
0.13134097, 0.16803309, -0.1724596, -0.1030807, -0.25199199,  
0.0466072, 0.38125145, 0.0306871, -0.19173753, 0.01305585,  
-0.07093169, -0.05639979, 0.07869941, 0.15101001, 0.02203472,  
0.0188807, -0.12816654, 0.04892636, 0.23726407, -0.10931174,  
-0.06033619, 0.2151441, -0.01290991, 0.10794014, 0.00208765,  
0.0523544, -0.03414371, 0.0557446, -0.1088978, -0.02816546,  
0.01456824, -0.07462301, -0.06795616, 0.1638117, -0.1352842,  
0.10086828, 0.00977207, 0.03836292, 0.02858941, -0.09801099,  
0.10640397, -0.03940248, 0.12043379, -0.27069992, 0.24780021,  
0.14584672, 0.13688208, 0.05027201, 0.09038105, 0.0349723,  
0.04752009, -0.02310574, -0.15151188, -0.00798576, 0.09137777,  
-0.01307906, 0.06952387, -0.03653277])
```



1.33



B인물의 사진1



Body
Embedding



```
array([-0.07440512, 0.13833548, 0.01550988, -0.04143689, -0.11708137,  
0.017141652, -0.09219746, -0.04101015, 0.12901564, -0.03510666,  
0.26017255, -0.01175268, -0.23214489, -0.10993981, -0.06433399,  
0.14533879, -0.18374404, -0.07009007, 0.01960303, 0.03927957,  
0.17917837, 0.10840175, 0.06728961, 0.01388871, -0.0921066,  
-0.32157087, -0.06875613, -0.11010363, 0.02105946, -0.09806888,  
0.095232023, -0.01714757, -0.16107245, -0.0485564, 0.05062422,  
0.04144309, -0.03346116, -0.03011878, 0.15263124, 0.01069992,  
-0.23673587, 0.05740198, 0.03090802, 0.23846291, 0.06076686,  
0.01160336, 0.00995683, -0.15661725, 0.09741502, -0.1173826,  
0.08130169, 0.15210505, 0.14222656, 0.02321066, 0.00600557,  
-0.07693202, -0.02959017, 0.15295519, -0.13042481, 0.03233673,  
0.1088061, -0.05199346, -0.01501178, -0.08011279, 0.17588341,  
0.02202863, -0.12124902, -0.26303217, 0.06608438, -0.128976,  
0.14779539, 0.1516934, -0.16154116, -0.1716671, -0.25228465,  
0.01865603, 0.36660314, 0.04856375, -0.18907131, 0.05604936,  
-0.05154709, -0.04398936, 0.08519959, 0.14640823, 0.00993889,  
0.02739849, -0.11102622, 0.01848426, 0.23328975, -0.1351117,  
-0.04641433, 0.22538319, -0.00492372, 0.10828383, 0.02823116,  
0.02502055, -0.02946196, 0.0702677, -0.09494975, -0.0334047,  
0.00996118, -0.08729131, -0.04657208, 0.09973253, -0.14260028,  
0.10422572, -0.02379006, 0.05201333, 0.02785442, -0.09933321,  
-0.09983464, -0.02418252, 0.13822252, -0.24259827, 0.2356335,  
0.12104575, 0.14091188, 0.07011457, 0.10868125, 0.04752614,  
0.02115094, -0.04581762, -0.23486597, -0.0105925, 0.11252803,  
-0.05433004, 0.10194337, -0.02596316])
```



1.27

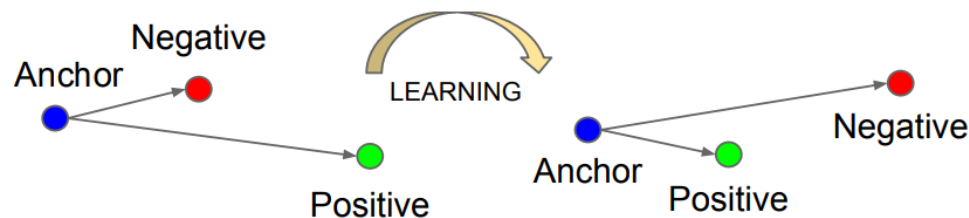
Method

Body Embedding

Name	Patch Size/Stride	Output Size
Conv 1	$3 \times 3/1$	$32 \times 128 \times 64$
Conv 2	$3 \times 3/1$	$32 \times 128 \times 64$
Max Pool 3	$3 \times 3/2$	$32 \times 64 \times 32$
Residual 4	$3 \times 3/1$	$32 \times 64 \times 32$
Residual 5	$3 \times 3/1$	$32 \times 64 \times 32$
Residual 6	$3 \times 3/2$	$64 \times 32 \times 16$
Residual 7	$3 \times 3/1$	$64 \times 32 \times 16$
Residual 8	$3 \times 3/2$	$128 \times 16 \times 8$
Residual 9	$3 \times 3/1$	$128 \times 16 \times 8$
Dense 10		128
Batch and ℓ_2 normalization		128

Triplet : 세 개의 데이터

- Anchor(x_i^a) : 기준 인물의 벡터
- Positive(x_i^p) : 기준과 같은 인물의 벡터
- Negative(x_i^n) : 기준과 다른 인물의 벡터



기준 인물과 같은 인물은 가깝도록, 기준 인물과 다른 인물은 멀도록

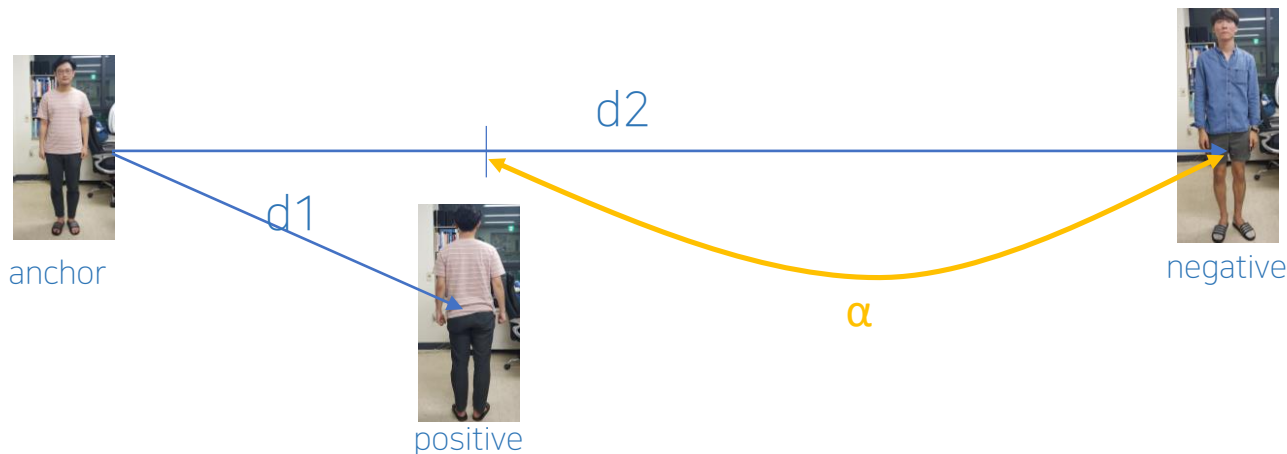
Method

Triplet Loss

$$\overset{d1}{\|f(x_i^a) - f(x_i^p)\|_2^2} + \alpha < \overset{d2}{\|f(x_i^a) - f(x_i^n)\|_2^2},$$

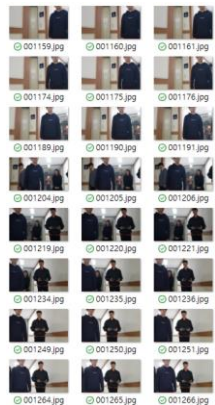
- x : 이미지
- $f(x)$: 임베딩 함수
- α : 마진
- x_i^a : 기준 인물(anchor) 이미지
- x_i^p : 기준과 같은 인물(positive)의 이미지
- x_i^n : 기준과 다른 인물(negative)의 이미지

Anchor와 Negative의 제곱거리가 Anchor와 Positive의 제곱거리보다 α 만큼 떨어져 있고 싶다!



Method

Triplet - mini batch



이미지셋

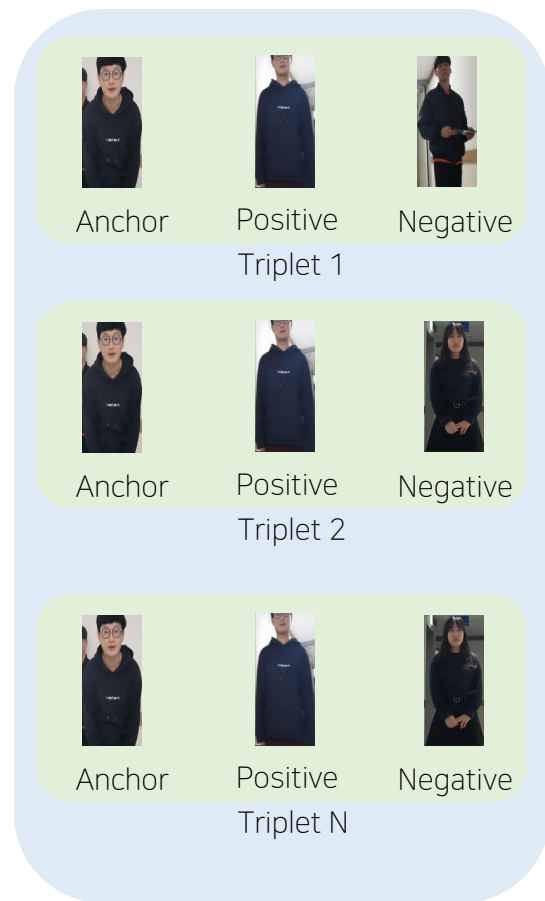
한 ID
40개 랜덤 샘플링

Negative faces
1960개 랜덤 샘플링



Random Sampling Batch

조합



Train Target

Method

Triplet – mini batch

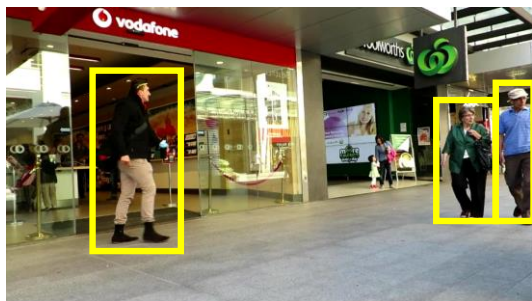


Method

Start Track

1. 탐지 결과를 사용해 트래킹 시작

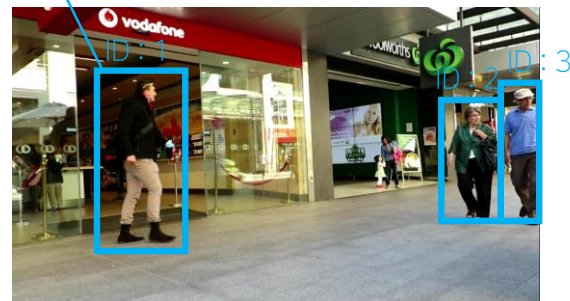
Frame 1



→
트래킹 시작

새 트랙 생성
`tracks += Track(id, feature)`

Frame 1

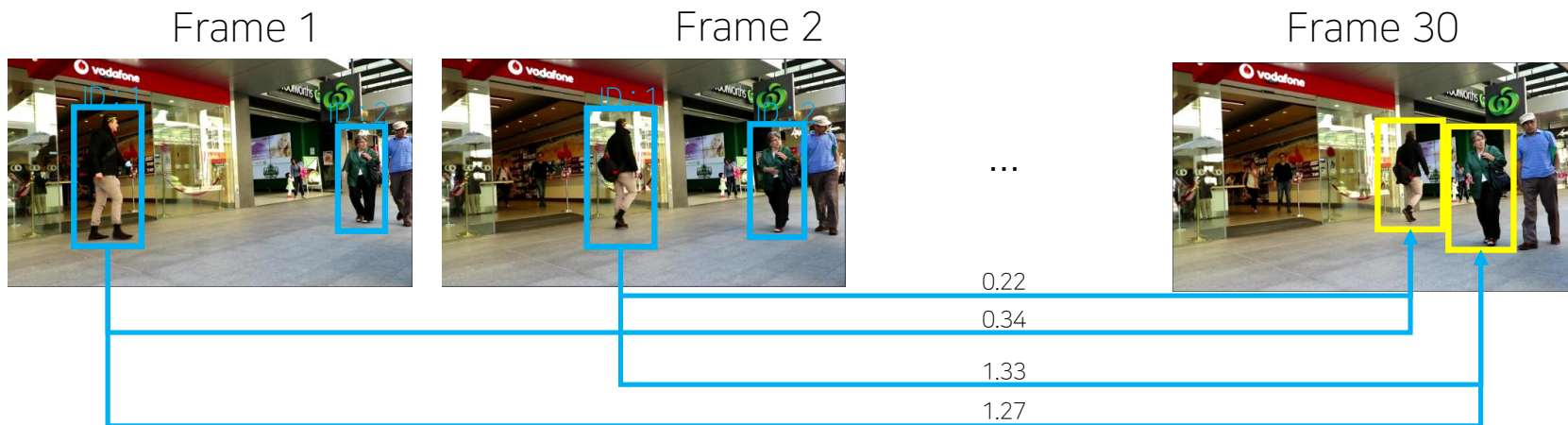


- 살아있는 트래킹
- 후보(탐지결과)

Method

Cascade Re-ID Matching

2. 후보와 이전 트래킹 결과를 Re-ID Model을 이용하여 매칭 - 순차적으로 과거 트래킹 피쳐와의 거리 계산



Body Feature의 L2 Distance가 Threshold 이하일 때 매칭
(과거 피쳐 일수록 엄격한 Threshold)

- 살아있는 트래킹
- 후보(칼만필터로 예측한 위치+탐지결과)

Method

Association - IOU

3. 후보와 2에서 매칭되지 않은 직전 트래킹 결과를 IOU를 이용하여 매칭



칼만필터로 예측한 위치와 탐지한 위치의 IOU가 Threshold 이상일 때 매칭

□ 살아있는 트래킹

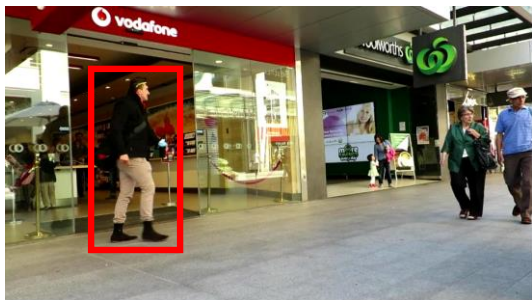
□ 후보(칼만필터로 예측한 위치+탐지결과)

Method

Association - New Track

4. 최종적으로 매칭되지 않은 탐지결과로 새 트래킹 시작

Frame 1

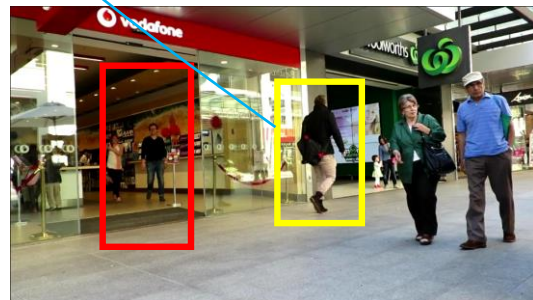


Frame 2



Tracking Object를 잃어버림

Frame 3



계속 Tracking Object를 잃어버림

새 트랙 생성
`tracks += Track(id, feature)`

- ❑ 잃어버린 트래킹 결과
- ❑ 후보(탐지결과)

Experiments

MOT Challenge



Ground Truth와 관심영역의 IOU가 0.5 이상일 때 True로 판단

MOTA : tracker 성능지표로 적합

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t}, \quad (1)$$

t : 프레임 인덱스

FN : 잘못탐지

FP : 놓친 객체

IDSW : id가 바뀐 횟수

GT : Ground Truth의 수

MOTP : detector 성능지표로 적합

$$\text{MOTP} = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t}, \quad (2)$$

t : 프레임 인덱스

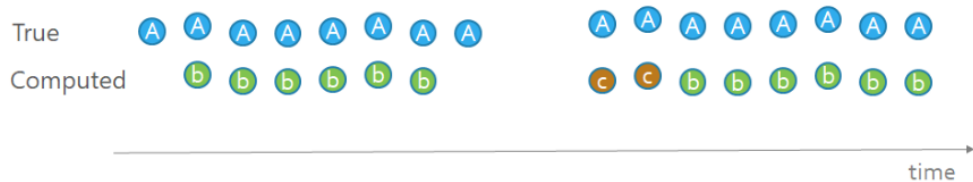
c_t : 탐지한 객체 항목의 수

$d_{t,i}$: 탐지한 객체와 Ground Truth를 비교한 IOU

Experiments

MOT Challenge

IDF1 : 얼마나 지속적으로 객체를 동일하게 판단하는가

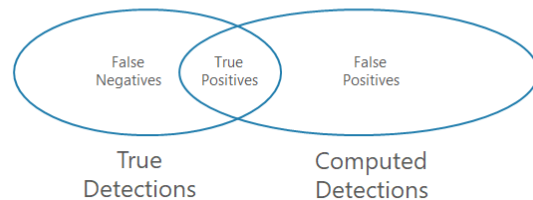


TP : 가장 많이 나온 ID의 개수 = 12

FP : 나머지 ID의 개수 = 2

FN : True - TP = 4

- ID Precision $P = \frac{TP}{TP+FP} = \frac{TP}{C}$
- ID Recall $R = \frac{TP}{TP+FN} = \frac{TP}{T}$
- F₁-score $F_1 = 2 \frac{PR}{P+R} = \frac{TP}{\frac{T+C}{2}}$



True Positive : True 이고, True라고 한 경우 -> 정답

False Positive : False 이지만, True라고 한 경우 -> 오답







True Negative : False 이고, False라고 한 경우 -> 정답

False Negative : True 이지만, False 라고 한 경우 -> 오답

Experiments

MOT Challenge

Test Set

	Sample	Name	FPS	Resolution	Length	Tracks	Boxes	Density	Description	Source	Ref.
1		MOT16-14	25	1920x1080	750 (00:30)	164	18483	24.6	Filmed from a bus on a busy intersection	link	[1]
2		MOT16-12	30	1920x1080	900 (00:30)	86	8295	9.2	Forward moving camera in a busy shopping mall	link	[1]
3		MOT16-08	30	1920x1080	625 (00:21)	63	16737	26.8	A crowded pedestrian street, stationary camera	link	[2]
4		MOT16-07	30	1920x1080	500 (00:17)	54	16322	32.6	A busy pedestrian street filmed at eye level by a moving camera	link	[2]
5		MOT16-06	14	640x480	1194 (01:25)	221	11538	9.7	Street scene from a moving platform	link	[3]
6		MOT16-03	30	1920x1080	1500 (00:50)	148	104556	69.7	Pedestrian street at night, elevated viewpoint	link	[1]
7		MOT16-01	30	1920x1080	450 (00:15)	23	6395	14.2	People walking around a large square.	link	[2]
	Total				5919 frm. (248 s.)	759	182326	30.8			

		MOTA ↑	MOTP ↑	MT ↑	ML ↓	ID ↓	FM ↓	FP ↓	FN ↓	Runtime ↑
KDNT [16]*	BATCH	68.2	79.4	41.0%	19.0%	933	1093	11479	45605	0.7 Hz
LMP_p [17]*	BATCH	71.0	80.2	46.9%	21.9%	434	587	7880	44564	0.5 Hz
MCMOT_HDM [18]	BATCH	62.4	78.3	31.5%	24.2%	1394	1318	9855	57257	35 Hz
NOMTwSDP16 [19]	BATCH	62.2	79.6	32.5%	31.1%	406	642	5119	63352	3 Hz
EAMTT [20]	ONLINE	52.5	78.8	19.0%	34.9%	910	1321	4407	81223	12 Hz
POI [16]*	ONLINE	66.1	79.5	34.0%	20.8%	805	3093	5061	55914	10 Hz
SORT [12]*	ONLINE	59.8	79.6	25.4%	22.7%	1423	1835	8698	63245	60 Hz
Deep SORT (Ours)*	ONLINE	61.4	79.1	32.8%	18.2%	781	2008	12852	56668	40 Hz



감사합니다.