

Mutfaktaki Günlük Davranışların Tanınması

Recognition of Daily Activities in Kitchen

Yazarlar Gizlenmiştir

Özetçe —Bu çalışmada, mutfak içindeki günlük davranışların giyilebilir kameralarla kaydedildiği Epic-Kitchens veri kümesine odaklanılarak, hareket ve nesne tanıma hedeflenmiştir. Literatürdeki en yeni yöntemler bu veri kümeye uyarlanmış, birden fazla yöntemi birleştirerek sonuçları iyileştirmek yönünde çalışmalar yapılmıştır. Ayrıca nesne ve hareketler arasındaki ilişkilerle, hareketlerin zamansal ilişkisinin performanstanı etkileri incelenmiştir.

Anahtar Kelimeler—*Nesne ve hareket tanıma, Davranış analizi.*

Abstract—In this study, we aim to identify the activities and objects in videos by focusing on the Epic-Kitchens data set where daily behaviors in the kitchen are recorded with wearable cameras. The state-of-the-art methods have been exploited for this data set and results are tried to be improved by combining multiple methods. In addition, effects of the temporal relationship of actions and the relationships between objects and actions have been experimented.

Keywords—*Action and object recognition, Activity analysis.*

I. GİRİŞ

Son yıllarda üretilen ve paylaşılan çok sayı ve çeşitlilikteki videonun erişilebilmesi için bu videoların analizi ve anlamlanırmaması önem taşımaktadır. Literatürde hareket tanıma için kontrollü veri kümeleri üzerinde yapılan çalışmalar, farklı ortamlarda farklı kişiler tarafından çekilen kontrolsüz videolarda yetersiz kalmaktadır. Yemek yapma gibi günlük hayatı ait olayların, veya bir işin nasıl yapıldığının yer aldığı videolarda hareket tanıma ilginç ve zor bir problem olmasının yanında, yalnız yaşayan yaşlı insanlara yardımcı olunması veya bir işin robotlara öğretilmesi açısından da faydalıdır.

Bu çalışmada günlük hayatı ait davranışların tanınması problemine odaklanılmış, mutfak içerisinde yapılan davranışlara ait nesne ve hareketleri tanımak hedeflenmiştir. Bu amaçla, kafa üzerine yerleştirilen kameralarla çekilmiş görüntülerden oluşan en büyük veri kümlesi olma özelliğinin yanında içeriği etiket çeşitliliği ve sayısı nedeniyle de en çok kullanılan veri kümelerinden olan Epic-Kitchens [1] veri kümlesi kullanılmıştır. Literatürde bu konuda yapılan çalışmaları geliştirebilmek adına hem nesne ve hareket tanımak için yeni modeller ve bu modellerin birlikte kullanımı alanında, hem de nesne-hareket ilişkisi ile hareketlerin zamansal olarak birbirini takip etme durumlarını göz önüne alarak varolan modelleri iyileştirme alanında yeni yöntemler önerilmiştir.

Aşağıda literatürdeki benzer çalışmaların özeti from mesinden sonra Epic-Kitchens veri kümeli tanıtılacak, ve önerilen yöntemlerin detayları verilecektir. Sonrasında her bir yöntem için sayısal ve görsel sonuçlar verilecek ve tartışılmaktır.

II. İLGİLİ ÇALIŞMALAR

Bir iş tarifinden adımların çıkartılması YouTube gibi kaynaklardaki sayıları gittikçe artan veriler nedeniyle son zamanlarda önem kazanmıştır. Yemek yapma gibi günlük hayatı ait olayların, ya da bir işin nasıl yapılacağını gösteren videoların toplandığı CrossTask [2] ve HowTo100M [3] gibi veri kümeleri ortaya çıkmıştır. Bu çalışmanın odaklandığı konu olan yemek videoları üzerinde aksiyon/aktivite tespiti alanında son yıllarda yapılmış çalışmalar bazları [4]–[6]’dır. Kuehne vd., (2015) Fisher Vector’leri kullanarak sabit kameralar ile alınan görüntülerden aktivite tespiti yapılmıştır. Li vd. [5] Egocentric kameralardan alınan görüntüler hareket (motion) ve obje (object) nitelikleri çıkartılmıştır.

Soran vd. [6] kişi-merkezli kameralar ile kayıt edilmiş yemek tarifi videoları üzerinde, atlamış ya da eksik yapılmış adımlar için uyarılar oluşturmayı amaçlamıştır. Bunu yapmak için öncelikle her tarif için, esnek ve sıralı bir çizge çıkartılmıştır. Daha sonra hem o ana kadar yapılmış aksiyonları hem de şu anda yapılmak üzere olan aksiyonu tespit etmek için Saklı Markov Modeli (Hidden Markov Model (HMM)) GIST nitelikleri ile kullanılmıştır. Atlanan eylemlerin ve maliyetlerinin çıkarımında, eylemlerin birbiriley zamansal ilişkileri, değişken sıralı çizge ile modellenmiştir. Bu yaklaşım, eylemlerin sebep-sonuç, birbirinin öenkoşulu olma gibi semantik ilişkisi ihmal edilmiştir.

Huang vd. [7], DIY türü öğretici videolarda bulunan pürüzler (noise), uyuşmayan ya da ses tanıma yöntemleriyle oluşturulmuş kayıtları sorunlarından yola çıkarak, bu tür sorunlara çözüm bulmak için kaynak çözümleme (“reference resolution”) olarak adlandırılan ifadelerin içerikle ilgili maddelelerle bağlantılılandırılması üzerine çalışılmışlardır. Bu çalışmanın sonucunda, örneğin ‘buz’ maddesinin ‘suyun dondurulması’ eyleminden üretmiş olduğu üzerine bir zamansal (temporal) bağlantı oluşturulabilmektedir. Kaynak çözümleme için öğreticisiz öğrenme yöntemini kullandıkları modelde, hem görsel ve hem linguistik model öğrenilmekte ve bu şekilde çeşitli belirsizliklere (ambiguities) karşı daha dayanıklı olunması sağlanmaktadır.

Zhou vd. [8] YouCook2 isimli web ortamından toplanmış

tarif videolarından oluşan yeni bir veri kümesi hazırlayarak bu veri kümesi üzerinde “Prosedür Böülütleme” adında yeni bir problem, bir tarif videosundaki bağımsız aşamaları tespit ederek böülütlemeye çalışmak, tanımlamışlar ve bunu çözmeye çalışmışlardır. Bu amaçla ProcNet isimli yeni bir ağ yapısı, video nitelikleri için ResNet metin nitelikleri için bi-directional LSTM kullanılmıştır, sunarak videolardaki prosedür segmentlerini tespit etmişlerdir.

III. VERİ KÜMESİ

Epic-Kitchens [1] veri kümesi birinci şahıs tarafından çekilen (egocentric) kamera görüntülerinden oluşmaktadır. Videolar dört farklı şehirde ve 32 adet mutfakta çekilmiştir. Tüm kayıtlar toplamda 55 saat sürmektedir. Videolar çerçevelere bölünmüştür ve veri kümesinde toplamda 11.5 milyon çerçeve yer almaktadır. Veri kümesi 125 hareket ve 331 nesne sınıfına ait etiket içermektedir. Hareket tanıma için 39.594 böülü, nesne algılama içinse 454.255 sınırlı kutu işaretlenmiştir. Epic-Kitchens veri kümesine ait örnek çerçeveler Şekil 1’de verilmiştir.



Şekil 1: Epic-Kitchens veri kümesine ait örnek çerçeveler.

Nesne tespiti için her bir çerçeveye ait etiket bilgisi mevcuttur. Etiket bilgisi içerisinde çerçevenin adı, çerçevede bulunan nesnelerin ait olduğu sınıf adı ve numarası, çerçevenin hangi mutfağa ait olduğu, çerçevenin ait olduğu video numarası ve çerçevedeki nesnelerin sınırlayıcı kutu bilgileri yer almaktadır. Şekil 2’de nesne algılama veri kümesine ait çerçeve bilgilerinden bir kesit gösterilmektedir.

```
noun_class,noun,participant_id,video_id,frame,bounding_boxes
20,bag,P01,P01_01,056581,"[(76, 1260, 462, 186)]"
20,bag,P01,P01_01,056611,"[(84, 1190, 446, 204)]"
20,bag,P01,P01_01,056641,"[(584, 936, 358, 268)]"
20,bag,P01,P01_01,056671,"[(472, 836, 412, 342)]"
```

Şekil 2: Nesne algılama için etiket bilgisi kesiti.

Hareket tanıma veri kümesine ait her bir çerçeve için de etiket bilgisi mevcuttur. Etiket bilgisi içerisinde hareketin başladığı ve tamamlandığı çerçevenin adı, hareketin başıldığı ve tamamlandığı zaman bilgisi, çerçevede bulunan hareketlerin ait olduğu sınıf adı ve numarası, çerçevede bulunan nesnelerin ait olduğu sınıf adı ve numarası, çerçevenin hangi mutfağa ait olduğu, çerçevenin ait olduğu video numarası gibi bilgiler yer almaktadır. Şekil 3’té hareket tanıma veri kümesine ait çerçeve bilgilerinden bir kesit gösterilmektedir.

Veri kümesinde bir video kesitindeki aktivite (activity), aktivitedeki nesne (noun) ve eylem (verb) birleşimi olarak ele alınmıştır. Dolayısıyla aktivite tespitinin doğru olabilmesi için

o kesitteki hem nesnenin hem de eylemin doğru bulunması gerekmektedir.

```
uid,participant_id,video_id,narration,start_timestamp,stop_timestamp,start_frame,stop_frame,verb,
verb_class,noun,noun_class,all_nouns,all_noun_classes
0,P01,P01_01,open door,00:00:00.14,00:00:03.37,8,202,open_2_door,8,['door'],[8]
1,P01,P01_01,turn on light,00:00:04.37,00:00:06.17,262,370,turn-on_12_light,113,['light'],[113]
2,P01,P01_01,close door,00:00:06.98,00:00:09.49,418,569,close_3_door,8,['door'],[8]
3,P01,P01_01,open fridge,00:00:12.77,00:00:13.99,766,839,open_2_fridge,10,['fridge'],[10]
```

Şekil 3: Aktivite tanıma için etiket bilgisi kesiti.

Veri kümesi çok büyük ve kapsamlı olmasına rağmen bazı hatalar da içermektedir. Örneğin bazı video kesitlerinde nesne olmasına rağmen nesne etiketi mevcut değildir. Bir başka sorun ise veri kümesindeki çerçeveler videoların kesitlere ayrılmış sonucu oluşturduğu için imgelerin bir kısmında netlik problemi vardır (Şekil 4).



Şekil 4: Epic-Kitchens veri kümesindeki net olmayan çerçeveler.

IV. DENEYLER

Epic-Kitchens veri kümesi üç farklı bilgisayarlı gözü problemi açısından ele alınmıştır. Bunlar sırası ile nesne algılama, eylem/hareket tanıma ve aktivite tanımadır. Deneylerde aktivite tespiti için kullanılan modeller her girdi için nesne ve eylem tahmini yapmaktadır. Eylem tespiti için yapılan doğruluk ölçümünde yalnızca modelin eylem sınıflandırma için yaptığı tahmin dikkate alınırken, aktivite tespiti için ise hem eylem hem de nesne tespiti için yaptığı sınıflandırma tahminleri dikkate alınmaktadır.

A. Nesne Algılama

İlk aşamada Epic-Kitchens veri kümesinde videolara ait çerçevelerdeki nesnelerin sınırlayıcı kutularla tanımlanması ve sınıflandırılması amaçlanmıştır. Bu amaç doğrultusunda Epic-Kitchens web sitesinde nesne algılama için oluşturulmuş, videolardaki kayık ve etiket olmayan görüntülerin temizlendiği çerçevelerden oluşan veri kümesi kullanılmıştır.

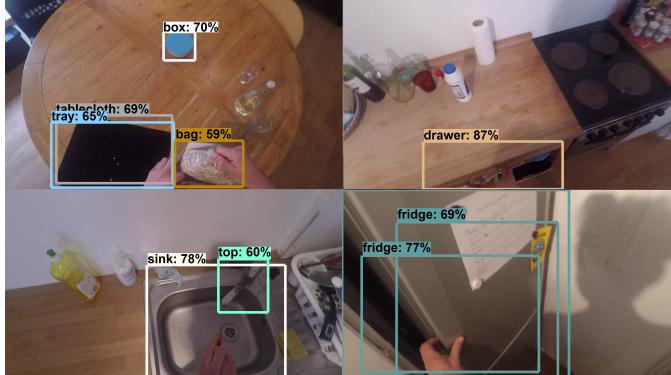
Nesne algılama yöntemi kaynak sıkıntısı sebebiyle veri kümelerinin küçük bir bölümünde gerçekleştirilmişdir. P01 numaralı mutfağa ait 01, 02, 05 ve 06 numaralı videolardan sırasıyla 3301, 1004, 2452 ve 984 adet imge eğitim veri kümesi olarak kullanılmıştır. Aynı mutfağa ait 03, 04, 08 ve 10 numaralı videolardan sırasıyla 238, 211, 198 ve 277 adet imge ise test için kullanılmıştır.

Nesne algılama metodunu olarak Faster RCNN [9] modeli Resnet-101 tabanıyla kullanılmıştır. Bu üç alt kümede alınan ortalama kesinlik (AP_{50}) değerleri Tablo I’de verilmiştir. Veri kümelerine ait yarışma sayfasında tüm veri kümesi kümeleri kullanılarak elde edilen en yüksek sonuç aynı metrik için

%34.18 olarak görülmektedir [10]. Şekil 5'te nesne algılama modeline ait örnek görsel sonuçlar paylaşılmıştır.

TABLO I: Faster RCNN modelinin nesne algılama sonuçları.

	P01-3	P01-4	P01-8	P01-10
Başarım (AP_{50})	18.57	12.29	10.41	9.95
Başarım (AP_{30})	23.85	16.52	14.97	15.21



Şekil 5: Nesne algılama modeline ait örnek sonuçlar.

B. Eylem/Hareket Tanıma

İkinci aşamada veri kümelerindeki videolardaki eylemlerin sınıflandırılması amaçlanmıştır. Bu amaçla 2B ve 3B Evrişimsel Sinir Ağlarını (ESA) içeren modeller kullanılmıştır. 2B ESA modelleri videoları çerçevelerine ayırarak işlerken, 3B ESA modelleri videoyu bütünsel olarak işlemektedir ve bu sayede zamansal ilişkileri daha iyi yakalayabilmektedir.

Mutfak aktivitelerini tanıyalımak amacıyla ilk olarak veri kümelerini sunanlar tarafından paylaşılan model [1] olan *Temporal Segment Networks (TSN)* [11] kullanılmıştır.

TSN modeline ek olarak, zamansal bağamları daha iyi yakalayabilmek adına bir 3B ESA modeli olan Inception-v1 I3D [12], Epic-Kitchens veri kümeleri üzerinde *finetune* edilmiştir. I3D modeli eğitilirken, modele aynı anda verilecek çerçeve sayısını ve seçilecek çerçevelerin pencere boyutunu belirlemek gerekmektedir. I3D [12] makalesinde çerçeve sayısı ve pencere boyutu 79 olarak seçilmiştir (79/79). Deneylerimizde bu değerlere ek olarak farklı pencere boyutları da kullanılmıştır, 79 çerçeve seçerken 200 çerçeve arasından rastgele bir çerçeveden başlayarak 79 sıralı çerçeve de seçtilik (79/200). Çerçeve ve pencere boyutu deneylerinde tüm veri kümeleri eğitme (%80), doğrulama (%10) ve test (%10) kümelerine ayrılarak kullanılmış ve modeller 5 epok boyunca, Adam optimizer ve 0.001 öğrenme oranı ile eğitilmiştir. Her iki deneye ait detaylı analiz sonuçları Tablo II'de verilmiştir.

TABLO II: 79/79 ve 79/200 deneyleri için doğrulama kümeleri üzerindeki hata ve başarım değerleri.

	Epok #1	Epok #2	Epok #3	Epok #4	Epok #5
Doğrulama Başarımı 79/79	0.11	0.13	0.23	0.24	0.26
Doğrulama Hatası 79/79	50.06	3.09	3.14	3.12	3.10
Doğrulama Başarısı 79/200	0.08	0.06	0.11	0.01	0.02
Doğrulama Hatası 79/200	3.40	935.03	29.84	2424.11	144.14

Sonuçlar incelendiğinde 200 pencere boyutu ile eğitilen modelin başarısının dalgalandığı ve başarılı bir şekilde eğitilemediği görülmektedir. Bu sebeple I3D modeli olarak, 79 çerçeve içinde 79 sıralı çerçeve ile eğitilen model kullanılmıştır.

TSN ve I3D modellerinin test kümelerindeki farklı video gruplarındaki ortalama başarımları Tablo III'te verilmiştir. Testlerde tüm test kümeleri yerine daha ufak bir alt kümeye kullanılmıştır, bu sebeple bizim aldığımız sonuçlar veri kümeleri bildirisindeki [1] sonuçlara göre az da olsa farklılık göstermektedir. Bu tablodan da görüleceği üzere ilk-5 metriğinde TSN, I3D modelinden daha başarılıdır. Sonraki aşamada, bu iki modelin çıktıları birleştirilerek başarımı artırmak hedeflenmiştir.

TABLO III: TSN ve I3D modellerinin ortalama test başarımları.

Model	Metrik	P01	P02	P03	P04	P05	P06	P07	P08	P10	Ort.
TSN	İlk 1	33.92	43.27	40.52	43.42	45.45	41.86	26.08	41.26	23.11	37.65
	İlk 5	83.92	81.26	79.29	78.94	86.36	95.34	60.86	76.19	56.52	77.19
I3D	İlk 1	19.80	29.10	23.34	12.63	8.82	22.32	9.17	34.92	23.34	20.38
	İlk 5	56.78	61.78	54.62	46.05	30.90	60.23	28.26	49.20	62.01	49.98

TSN ve I3D sonuçlarını birleştirmek için aşağıdaki gibi bir formül kullanılmıştır:

$$Model = \alpha * I3D + (1 - \alpha) * TSN$$

Tablo IV'te farklı α parametreleri için elde edilen ilk-1 ve ilk-5 eylem tanıma başarımları verilmiştir. Sonuçlar incelendiğinde α parametresi 0.1 seçildiğinde her iki başarım metriği için de sonuçların yükseldiği görülmektedir. TSN modelinin ortalama başarımı ilk-1 metriği için %37.5'ten %39.0'a, ilk-5 metriği için ise %77.2'den %79.1'e yükselmiştir. Bu da zamansal bilginin aktivite tespitindeki önemini göstermektedir.

TABLO IV: Farklı α parametreleri için ortalama model başarımları.

Model	$\alpha = 0.1$		$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 0.75$		$\alpha = 0.9$	
	İlk 1	İlk 5	İlk 1	İlk 5	İlk 1	İlk 5	İlk 1	İlk 5	İlk 1	İlk 5
	39.02	79.12	36.94	79.83	34.85	77.63	29.67	75.44	23.71	68.57

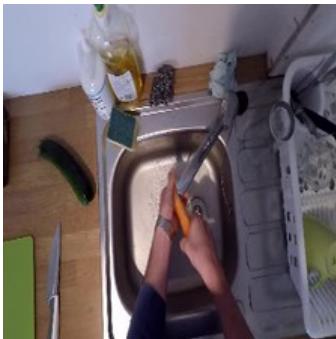
C. Nesne-Eylem İlişkilerinin Kullanılması

Epic-Kitchens veri kümelerinin içeriği aktiviteler içerisinde iki temel yapı taşı bulunmaktadır: eylem ve eylemden etkilenen nesne. Üzerinde çalıştığımız model olan TSN, aktivite tespiti yaparken aktivitelerin bileşenlerini ayrı ayrı tahmin edip

sınıflandırmaya çalışmaktadır. Yani modele sınıflandırılması için bir aktivite verildiğinde, model aktiviteyi oluşturan eylem ve nesne için farklı güven skorları üretmektedir. Aktivitenin final skoru bu iki çıktı birleştirilerek elde edilmektedir. Bu çalışmada, temel metodun aldığı sonuçlar iki çözüm önerisiyle geliştirilmeye çalışılmıştır.

İlk olarak eylemi oluşturan nesne ve eylemin birbirileyle olan ilişkisini kullanarak temel modelin aktivite tahminleri güncellenmeye çalışılmıştır. Bu ilişkiye entegre etmemizin sebebi; aktivite tespiti yapılırken nesne ve eylem ayrı ayrı tahmin edildiğinde uyumsuz durumların (Örneğin Şekil 6), eğitim ve risinden öğrenilen nesne-eylem ilişkisiyle düzeltilemesidir. Test kümesinde elde ettigimiz sonuçları incelediğimizde karşımıza doğru tespit edilen bir eylem için o hareketle bulunma ihtimali çok düşük olan yanlış tahmin edilen nesne örnekleri çıkmıştır.

Aktivite tespitinin doğruluğunu artıracak ikinci çözüm önerimiz ise eylemlerin içerdiği hareketlerin birbirleri ile ilişkisini kullanmaktadır. Epic-Kitchens veri kümesi temel olarak bir insanın mutfak içindeki günlük eylemlerini barındırmaktadır. Dolayısıyla bu aktiviteler belirli bir akışa göre olmaktadır. Bu bilgi göz önüne alındığında, modelimizin tespit ettiği eylemlerin birbirinden sonra gelmelerinin rastgele değil bir örüntüye göre sıralandığı varsayımda bulunulmuş ve verinin eğitim setinden bu örüntüler elde ederek, test esnasında modelin tahminlerini güncelleyip doğruluğu artırmaya çalışılmıştır.



Şekil 6: Model görsel için çöpü yıkama tahmininde bulunmuştur. Eylem doğru tespit edilmiş olmasına rağmen nesne hatalı bulunduğu için hatalı bir aktivite tespiti olarak işaretlenmiştir.

Önerdiğimiz çözümler için TSN modeli dayanak olarak seçilmiştir. Eylem ve nesne ilişkilerini eğitim veri kümesinden elde etmek amacıyla, satırları hareketleri, sütunları ise nesneleri temsil eden iki boyutlu bir matris oluşturulmuştur. Bu dizinin elemanları eylemlerin nesneler ile görülmeye sayılarını belirtmektedir.

Bu matrisi incelediğimizde, örneğin alma (take) eyleminin en çok bıçak, kaşık, tabak gibi nesneler ile sıkılıkla kullanılan gözlemlenmiştir. Bunun anlamı, alma eylemi modelimiz tarafından tahmin edildiğinde bu eylemle geçmesi olası bıçak, kaşık, tabak gibi nesnelerin güven skorunu artırması, aksine alma eylemiyle alakasız olan örneğin kapı nesnesinin güven skorunu düşürmesidir. Bu yöntemin, ilk-1 başarım metriği için %0.8 ve ilk-5 metriği için %1.7'lik bir artış sağladığı tespit edilmiştir.

Eylemler arasındaki ilişkileri elde etmek amacıyla, satırları mevcut hareketi, sütunları ise mevcut hareketten sonra hangi

hareketin gelebileceğini temsil eden bir matris oluşturulmuştur. Bu dizinin elemanları bir eylemden sonra gelebilecek diğer eylemlerin görülmeye sayılarını belirtmektedir.

Bu iki boyutlu matrisi kullanarak modelin tahminleri güncellenmiştir. Ancak önerdiğimiz yöntemin sonuçları kötü yönde etkilediği ve başarımı düşürdüğü gözlenmiştir. Bunun sebebi olarak, veri kümesinin yemek hazırlama dışında birbirileyle ilişkisi olmayan bazı eylemlerin ard arda gelmesi ve dolayısıyla eğitim kümesinden hareketler arasında sağlıklı bir ilişkinin kurulamaması olduğu düşünülmüştür. Epic-Kitchens yerine daha görev temelli bir veri kümesinde bu çözümün işe yarayabileceği düşünülmektedir.

V. SONUÇ

Bu çalışmada Epic-Kitchens veri kümesi üzerinde nesne algılama, eylem tanıma ve aktivite tespiti problemleri üzerine yoğunlaşılmıştır. Mevcut yöntemlere ek olarak, eylem tespitinin başarımını artırmak için model birleştirme ile aktivite tespitinin başarımını artırmak için nesne/eylem ilişkisi ve eylem akış bilgileri kullanılmıştır. Bunlardan aktivite akış bilgisi kullanmanın başarımı düşürdüğü, diğer yöntemlerin ise başarımı artırdığı gözlenmiştir.

KAYNAKLAR

- [1] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, “Scaling egocentric vision: The epic-kitchens dataset,” in *European Conference on Computer Vision (ECCV)*, 2018.
- [2] J.-B. A. I. L. D. F. Ramazan Gokberk Cinbis, Dimitri Zhukov, “Cross-task weakly supervised learning from instructional videos,” 2019.
- [3] J.-B. A. M. T. I. L. J. S. Antoine Miech, Dimitri Zhukov1, “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips,” 2019.
- [4] H. Kuehne and T. Serre, “Cooking in the kitchen: A generative approach to the recognition, parsing and segmentation of human daily activities,” *CoRR*, 2015.
- [5] Y. Li, Z. Ye, and J. M. Rehg, “Delving into egocentric actions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 287–295, 2015.
- [6] B. Soran, A. Farhadi, and L. Shapiro, “Generating notifications for missing actions: Don’t forget to turn the lights off!,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4669–4677, 2015.
- [7] D.-A. Huang, J. J. Lim, L. Fei-Fei, and J. Carlos Niebles, “Unsupervised visual-linguistic reference resolution in instructional videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2183–2192, 2017.
- [8] L. Zhou, C. Xu, and J. J. Corso, “Towards automatic learning of procedures from web instructional videos,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [9] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, pp. 91–99, 2015.
- [10] “Epic-kitchens object detection competition.” <https://competitions.codalab.org/competitions/20111results>.
- [11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *European conference on computer vision*, pp. 20–36, Springer, 2016.
- [12] A. Z. Joao Carreira, “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset,” 2018.