Bilgi Damıtma ile Robot-Nesne Etkileşim Hatalarını Tahminleme
Robot-Object Manipulation Failure Anticipation using Knowledge Distillation

Tuğçe TEMEL, Arda İNCEOĞLU, Sanem SARIEL

Yapay Zeka ve Robotik Laboratuvarı (AIR Lab), İstanbul Teknik Üniversitesi
İstanbul, Turkiye
{temel21, inceoglua, sariel}@itu.edu.tr

***Özetçe* —Otonom bir hizmet robotunun çevresiyle emniyetli bir şekilde etkileşim kurabilmesi gerekir. Ancak, algılama hataları, nesne etkileşimi yanlışlıkları veya beklenmedik dış etkenler gibi çeşitli belirsizlikler nedeniyle nesne etkileşimi sırasında hatalar meydana gelebilir. Mevcut araştırmalar genellikle robot hatalarının tespiti ve sınıflandırması problemine odaklanmışken, bu çalışma bu tür hataların önceden tahminine odaklanmaktadır. Temel varsayım, bir hata yeterince erken öngörülebilirse, önleyici eylemlerin alınabileceğidir. Bu amaçla, bu bildiride yeni bir bilgi damıtma tabanlı tahminleme mimarisi sunulmuştur. Önerilen mimari, video dönüştürücülerin gücünden faydalanarak RGB, derinlik ve optik akış verilerini işleyebilen çok kipli bir sensör füzyon ağı içermektedir. Yaklaşımımızın başarımı FAILURE adlı gerçek dünya robot nesne etkileşimi veri kümesi üzerinde değerlendirilmiştir. Deneysel sonuçlar, önerdiğimiz mimarinin robot yürütme hatalarını gerçekleşmeden 1 saniye öncesinde tahmin etme konusunda %82.12 F1 skoru elde ederek bu konudaki etkinliğini ortaya koymaktadır.**

***Anahtar Kelimeler*—Hata tahminleme, Bilgi Damıtma, Dönüştürücü**

***Abstract*—An autonomous service robot should be able to safely interact with its environment. However, failures can occur during manipulation execution due to various uncertainties such as perception errors, manipulation inaccuracies, or unforeseen external events. While existing research has primarily focused on the detection and classification of robot failures, this work focuses on anticipation of such failures. The premise is that if a failure can be anticipated early enough, prevention actions can be taken. To this end, we introduce a novel knowledge distillation-based anticipation framework. Our framework leverages the power of video transformers and incorporates a multimodal sensor fusion network capable of processing RGB, depth, and optical flow data. We evaluate the success of our approach using a real-world robot manipulation dataset named FAILURE. Experimental results demonstrate that our proposed framework achieves an 82.12% F1 score, showcasing its efficacy in anticipating robot execution failures up to 1 second in advance.**

***Keywords*—Failure Anticipation, Knowledge Distillation, Transformers,**

979-8-3503-8896-1/24/$31.00 ©2024 IEEE

I. GİRİŞ

Robotlar ev ve ofis gibi sosyal ortamlarda insanlarla ve nesnelerle ortak bir etkileşim içerisindedir. Bu gibi düzensiz (unstructured) ortamlarda kullanılan robotlar, çeşitli etkileşim becerilerine ihtiyaç duyarlar [1]. İnsan-robot veya robot-nesne etkileşimi sırasında robot ya da çevresel faktörler (örn., dış olaylar) kaynaklı istenmeyen durumlar oluşabilir. İstenmeyen durumlar; etkileşim parametrelerinin hatalı tahmini, dünyanın içsel temsilinin eksik veya yanlış tanımlanması nedeniyle meydana gelebilir. Bu gibi durumlar robot-nesne etkileşiminde güvenlik ve emniyet bakımından endişelere neden olmaktadır.

Emniyetin sağlanması için robot nesne etkileşim hatalarının tespiti [27] ve sınıflandırılması [3], [8], [9] problemleri sıklıkla ele alınmıştır. Fakat hataların sadece tespit edilip sınıflandırılması, emniyetli çalışma için yeterli değildir. Hataların meydana gelmeden önce tahminlenerek (anticipation) önleyici aksiyonlar ile meydana gelmesini engellemek daha etkili bir çözüm olacaktır. Bu sayede hatanın verebileceği potansiyel zarar en aza indirgenebilinir.

Bu bağlamda, yakın zamandaki bir çalışmada, robotun derin pekiştirmeli öğrenme yöntemleri ile hata önleyici eylemleri öğrenmesi sağlanmıştır [10]. Fakat bu çalışmada, emniyeti bozan risklerin tahminlenmesi için kural bazlı yöntemler kullanılmıştır. Önerilen yöntemde ise bu amaca hizmet etmek üzere tahminlemenin uzman bilgisi olmadan ve önceden belirlenmiş öznitelikler kullanılmadan,gerçek dünya verileri ile yapılması hedeflenmiştir.

Tanımlara bakıldığında; erken tespit (early detection), bir olayın mümkün olduğunca erken tespit edilmesi, tahminleme (anticipation) ise bir olayın gerçekleşmeden önce tahmin edilmesidir [11]. Literatürde, ele alınan hata tahminleme problemine benzer olarak insan aktivite tahminleme problemi üzerinde çalışmalar mevcuttur [12]. Örneğin, [13] gecikmiş tahminler için kaybı üstel artıran bir çapraz entropi kaybı (exponential loss) kullanırken, başka bir çalışmada [14] çapraz entropi kaybı ile sıralama kaybı (ranking loss) birleştirilir. Bir başka çalışmada [15], Rolling-Unrolling LSTM yöntemi kullanılır. Bu yöntemde Rolling LSTM, tarihsel bağlamı bir gizli vektöre kodlar ve Unrolling LSTM bu gizli vektörü

çözmek için kullanır. Bilgi damıtma ise, büyük ve karmaşık bir modelin (Öğretmen model) bilgisini, daha küçük ve hafif bir modele (Öğrenci model) aktarma sürecidir. Bu süreç, genellikle öğrenci modelin performansını artırmak veya onu gerçek zamanlı uygulamalara uygun hale getirmek amacıyla gerçekleştirilir [1618]. Bilgi damıtma yapısı insan aktivitelerinin tahmin edilmesinde de kullanılmaktadır [19], [20]. Bu çalışmada bilgi damıtma, literatürdeki yöntemlerden farklı olarak robot nesne etkileşimi hatalarını tahminleme için tekrar yorumlanmış ve çok kipli bir mimari kapsamında kullanım şekli BÖLÜM-II-B2 kısmında ayrıntılı olarak anlatılmıştır.

Gerçek dünya verileri ile video üzerinden robot-nesne hataların tahminlenmesi için sensör verilerinin uzamsal-zamansal ilişkinin analizi gerekmektedir. Son yıllarda video dönüştürücüler bu analizleri kullanarak, bilgisayarlı görüde önemli ilerleme kaydedilmesini sağlamıştır. Uzamsal ve zamansal bağımlılıkları etkili bir şekilde yakalamak için, video dönüştürücü tüm kare dizilerini aynı anda işler. Öz-odaklanma (self-attention) mekanizmalarıyla karmaşık etkileşimleri modellemede başarılı performans sergilerler, bu da onları video işleme gibi görevler için ideal hale getirir. Bu sebeplerden ötürü, temel model olarak ViViT [21] kullanılmıştır.

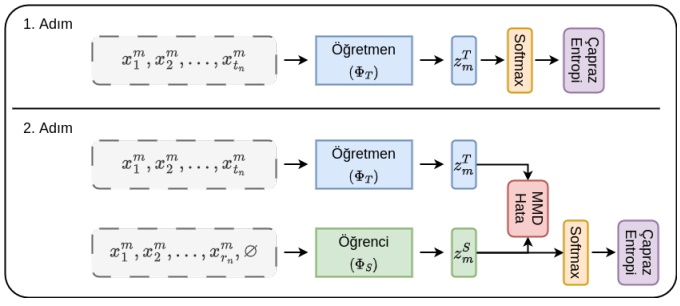
Bu çalışmanın katkıları özetle şunlardır: (i) optik akış (optical flow) kipi ile genişletilmiş FAILURE [2] veri kümesinin oluşturulması, (ii) odaklanma tabanlı bir çok kipli dönüştürücü (transformer) mimarisinin kullanılması, (iii) bilgi damıtma yolu ile elde edilmiş hata tahminleme modelinin geliştirilmesidir.

II. Bilgi Damıtma Tabanlı Hata Tahminleme

Bu çalışmada bilgi damıtma tabanlı bir hata tahminleme mimarisi sunulmuştur (Şekil 1). Önerilen mimaride *Öğretmen* ve *Öğrenci* ağları bulunmaktadır. *Öğretmen* ağ öğrenme sürecine yardımcı olurken, hata tahminleme *Öğrenci* ağ tarafından yapılmaktadır.

A. Problem Tanımı

Robotun yürütme esnasındaki hatalarının ortak örneklemini tanımlamak için, çoklu sensör kipleri (RGB, derinlik ve optik akış bilgileri) kullanarak sahnedeki değişiklikleri dikkatle gözlemlemek gerekir. Bu çalışma kapsamında, hatanın gerçekleştiği görüntü çerçevesi bilgisi olmaksızın hataları tahminleme görevi bir sınıflandırma problemi olarak kavramsallaştırılmıştır. $M \in \{1, 2, 3 \ldots\}$, kip kümesini temsil eder ve $m \in M$ kip tipini belirtir. D^*, $|D^*| = N$ olmak üzere, farklı kip gözlem dizilerini içeren veri kümesini temsil eder. Bir gözlem $x^m_{t_m,i}$ ile gösterilir. t_m, m kipinin zaman indisini temsil eder. i ise kayıt indeksidir. Son olarak, $y \in \{hatali, basarili\}$ olmak üzere sınıflandırma etiketlerini belirtir.



Şekil 1: Bilgi damıtma tabanlı hata sezme mimarisi. İlk adımda Öğretmen (Φ_T) ağı eğitildikten sonra ikinci adımda, eğitilmiş Öğretmen ağının ağırlıkları dondurularak, yalnızca Öğrenci (Φ_S) ağı eğitilir. Öğretmen ağının girdi karelerinin (x_m) örneklenmesi sırasında videoya ait tüm kareler kullanılırken, öğrenci ağının girdisinde son kare yerine boş kare ($\varnothing$) verilir. Bu sayede ileriye yönelik tahmin yapmayı öğrenmesi hedeflenir.

$$D^* = \{(x^m_{1,i} \ldots x^m_{t_m,i}))^M_{m=1}, y_i\}^N_{i=1} \quad (1)$$

Amaç, tüm video verisi ile eğitilmiş ve hata tespiti için tasarlanmış $\Phi_T(.)$'dan yürütme esnasında nesne etkileşiminin tüm ayırt edici özelliklerini çıkararak, bu özellikleri her nesne etkileşimi tipinin sadece hata olmadan önceki kısımlarının bilgisine sahip olan $\Phi_S(.)$ fonksiyonuna damıtmaktır.

B. Sinir Ağı Mimarisi

1) Görü Dönüştürücü: Görüntülerdeki öznitelikleri çıkarmak için kullanılan temel mimari ViViT [21] (Video Vision Transformer) görü dönüştürücüdür. ViViT modeli için görüntü girdi çerçeve sayısı (t) 8, görüntü çerçevesi boyutu (h, w) 224 x 224, görüntü çerçevesinden çıkarılan dilimin boyutu ($patch_h$, $patch_w$) 32 x 32, zamansal dilim boyutu ($patch_t$) 4, boyut parametresi 256, derinlik parametresi 6, çok başlıklı özodaklanma (multi head self-attention) boyutu 8 olarak seçilmiştir.

2) Bilgi Damıtma: Bu çalışmada önerilen, bilgi damıtma tekniğinin uygulanması yoluyla bir ağın sahip olup (*Öğretmen* Ağı) diğer ağın (*Öğrenci* Ağı) sahip olmadığı bilgileri birinden diğerine aktarmaktır. Literatürdeki çalışmalardan farklı olarak *Öğretmen* ve *Öğrenci* ağlarının parametre sayıları birbirine denktir.

Öğretmen Ağı: Öğretmen Ağı hata tespiti için eğitilmiş ve sürece dahil olan tüm aksiyon video çerçeveleri hakkında bilgi sahibidir. Öğretmen modeli aşağıdaki gibi formüle edilen Φ_T sembolü ile temsil edilmiştir. Formülde ϕ_m farklı kiplere karşılık gelen sinir ağıdır.

$$\hat{y}_{t_m,i} = \Phi_T(\phi_1(x^{(1)}_{1,1}, \ldots, x^{(1)}_{t_1,i}) \oplus \cdots \oplus \phi_m(x^{(m)}_{1,1}, \ldots, x^{(m)}_{t_m,i})) \quad (2)$$

Öğrenci Ağı: Öğrenci ağının temel amacı hataları önceden tahminlemektir. Bu hedef sembolik olarak Φ_S ile temsil edilir. D veri setindeki bir gözlem $x^m_{r_m,i}$ ile ifade edilir. Zamana bağlı t_m parametresinin alt kümesi olan r, $r \in \{1, 2, \ldots, (t-1)_m\}^M_{m=1}$ olarak tanımlanır. Φ_s ile temsil edilen öğrenci için hata tahminleme görevi, r_m değişkenine bağlı olarak ile Denklem 3'te formüle edilmiştir.

$$\hat{y}_{t_m,i} = \Phi_S(\phi_1(x^{(1)}_{1,1}, \ldots, x^{(1)}_{r_1,i}) \oplus \cdots \oplus \phi_m(x^{(m)}_{1,1}, \ldots, x^{(m)}_{r_m,i})) \quad (3)$$

C. Eğitim Süreci

Eğitim stratejisi iki ana adımdan meydana gelmektedir. İlk adımda yalnızca *Öğretmen* ağı eğitilir. İkinci adımda eğitilen *Öğretmen* ağı dondurularak *Öğrenci* ağı eğitilir. *Öğrenci* ağının eğitimi sırasında kullanılan hata fonksiyonu 2 adet hata(loss) fonksiyonunun bileşiminden oluşmaktadır. Bunlardan bir tanesi öğrenci ağının hata tahminleme hatasını diğeri

ise öğretmenden-öğrenci ağındaki bilgi damıtma hatasını ölçmek için kullanılmıştır. Toplam hata fonksiyonumuz denklem 4'teki gibi formüle edilmiştir.

$$L_{toplam} = \alpha.L_{TS}(z_m^T, z_m^S) + \beta.L_C(y, \hat{y}) \quad (4)$$

α ve β parametreleri ağ için öğrenilebilir parametreler olarak tanımlanmış olup değerleri eğitim aşamasında ağ tarafından belirlenmektedir. L_C öğrenci ağının tahmin hatasıdır ve kategorik çapraz entropi (CCE) ile hesaplanır. y_m, m kipinin video kayıt indeksinin gerçek sınıf etiketi ve $\hat{y}_m$ aynı indis ve kipteki videonun öğrenci hata tahminlemesidir. Bilgi damıtma hatasını temsil eden L_{TS} maksimum ortalama farklılık (MMD) hata fonksiyonu kullanılarak hesaplanır ve Denklem 5'teki gibi formüle edilmiştir.

$$L_{TS} = L_{MMD}(z_{RGB}^T, z_{RGB}^S) + L_{MMD}(z_D^T, z_D^S) + L_{MMD}(z_F^T, z_F^S) \quad (5)$$

Öğretmen-Öğrenci eğitimi çiftinde öğretmenin öğrenciden bağımsız bir şekilde önce eğitildiği ve öğrenci ağı eğitimi sırasında tekrardan eğitime dahil edilmeyip sadece bilgisinin damıtıldığı durum çevrimdışı damıtma olarak tanımlanır. Çalışmamızda öğretmen-öğrenci çifti eğitiminde, diğer sınıflandırma problemlerinde de benimsendiği gibi çevrimdışı damıtma (offline distillation) benimsenmiştir [2224]. Önce öğretmen ağı eğitilmiştir ve ardından sadece öğrenciye, önceden öğrendiği bilgileri damıtarak rehberlik etmiştir.

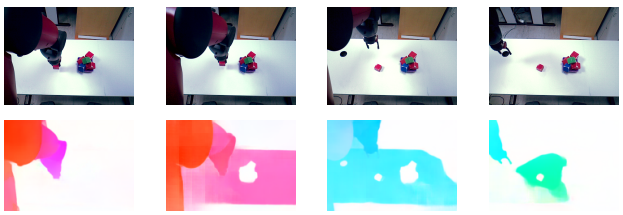
III. Deneyler

A. Veri Kümesi

Bu çalışmada FAILURE [3] veri kümesi kullanılmıştır. Bu veri kümesi, Baxter robotu kullanılarak elde edilmiş gerçek dünya RGB, derinlik ve ses içeren 324 nesne etkileşimi video kaydından oluşturulmuştur. Bu nesne etkileşimleri *dökme, itme, kaba-yerleştirme, üst-üste-yerleştirme ve tutma&bırakma* olarak adlandırılmıştır.

FAILURE veri kümesindeki kiplere ek olarak optik akış (optical flow) kipi de eklenerek veri seti zenginleştirilmiştir. Optik akış, bir görüntüdeki parlaklık desenlerinin hareketinin görünür hızlarını ifade eden bir dağılımdır. Bu hareket, hareket kaynağı olan unsur veya edilgen olarak bundan etkilenen nesnelerden göreceli olarak kaynaklanabilir [25]. Bu nedenle, optik akış, etken ve edilgen olarak harekete neden olan nesnelerin uzamsal düzenlemesi ve bu düzenlemenin değişim hızı hakkında önemli bilgiler sağlar [26].

Hata tahminleme problemi için ortamdaki nesnelerin yeri, uzaklığı ve hareket durumları bilgisine sahip olmak hatanın imzasını çıkarabilmek için önemli bir kriterdir. Bu sebeple ham RGB verisine ek olarak derinlik ve optik akış verisinin birlikte kullanılması önerilmiştir. Optik akış verisi, RGB verileri üzerinden FlowNet2 [27] kullanılarak elde edilmiş ve veri kümesine eklenmiştir. Şekil 2'de RGB ve optik akış kare çiftleri sunulmuştur.



Şekil 2: Nesne etkileşimi sırasında robotun kafa kamerasından alınan RGB kareler ve hesaplanan optik akış eşleri

TABLO I: Öğretmen Ağın Başarımı

	Hassasiyet	Duyarlılık	F1 Skoru
RGB	62.63	62.16	62.39
D (Derinlik)	64.97	60.89	62.86
F (Optik Akış)	64.57	62.16	63.34
RGB-F	67.56	67.56	67.56
RGB-D	70.43	67.56	68.97
RGB-D-F	**76.05**	**75.67**	**75.86**

TABLO II: Öğrenci Ağı Başarımı (Hata Tahminleme)

	1-Çerçeve (1 sn)			2-Çerçeve (2 sn)		
	Hassasiyet	Duyarlılık	F1 Skoru	Hassasiyet	Duyarlılık	F1 Skoru
RGB	73.05	72.97	72.98	72.64	62.16	66.99
D	68.26	68.42	68.34	66.84	65.78	66.31
RGB-D	75.85	75.67	75.76	70.19	70.27	70.23
RGB-F	79.41	78.37	78.89	67.48	67.52	67.52
RGB-D-F	**83.19**	**81.08**	**82.12**	**81.27**	**78.37**	**79.79**

Veri Ön İşleme: Robot, nesne etkileşimi yörüngesini çevrimiçi olarak planlayıp yürüttüğünden dolayı kayıtların uzunluğu, nesne etkileşimi tipi ve planlanan yörüngeye bağlı olarak değişiklik gösterir. Veri kümesindeki her bir kayıt 1 FPS örneklenerek sıralı görüntü çerçevesi serilerine dönüştürülür. Veri kümesindeki videolar rastgele örneklenmiş 8 çerçeve ile temsil edilmektedir. *Öğretmen* ağı, 8 çerçevenin tamamını gözlemleyebilirken, *Öğrenci* ağı için seçilebilecek son çerçeve hatadan önceki olacak şekilde kısıtlanmıştır. Son olarak, tüm çerçeveler masa düzlemine karşılık gelen 224 × 224 piksel boyutunda kırpılmıştır.

B. Deney Düzeni

Nicel değerlendirme için, genişletilmiş FAILURE veri kümesi eğitim (%70), doğrulama (%10) ve test kümeleri (%20) olarak bölünmüştür. Tüm ağ ağırlıkları rastgele başlatılmaktadır ve Adam optimizasyonu kullanılarak 250 adım boyunca eğitilmektedir, öğrenme hızı $1e-5$ olarak belirlenmiştir. En iyi model doğrulama kümesi üzerinde erken durdurma stratejisi ile belirlenmektedir. Seçilen en iyi modellerle elde edilen test skorları aşağıdaki bölümlerde raporlanmıştır. Aşırı öğrenmenin (overfitting) önlenmesi için, veriye renk artırma ve rastgele çevirme uygulanır. Örneğin, bir dizideki tüm görüntülerin parlaklık, kontrast, doygunluk ve ton değerleri %20 olasılıkla rastgele değiştirilir. Benzer şekilde, her görüntü dizisi dikey olarak %50 olasılıkla çevrilir.

C. Sayısal Sonuçlar

Önerilen mimari, hem tekil kipler için hem de çok kipli sensör füzyonu olarak eğitilmiştir. Çok kipli olan mimaride her bir kip için dönüştürücü ağının bağımsız bir kopyası oluşturulur ve kipler geç füzyon (late fusion) ile birleştirilir.

Tablo I'de *Öğretmen* ağının farklı sensör kipleri ve bunların füzyonu sonucunda elde edilen sonuçlar sunulmuştur. *Öğretmen* ağının girdisini oluşturan çerçeveler, videonun tamamını kapsadığı için hata tespit problemi olarak da nitelendirilebilir. Tablo analiz edildiğinde farklı kiplerin birleştirilmiş olduğu bir

ağın tekil kip ile eğitilen ağdan, hata tespiti bakımından daha iyi performans elde ettiği görülmüştür. En iyi performans tüm kiplerin füzyonu (RGB-D-F) ile elde edilmiş ve bu performans test aşamasında F1 skoru bazında $\%76.06$ olarak ölçülmüştür. Tablo I göz önüne alındığında *hata tahminleme problemi* bazında çoklu kip kullanımının daha iyi performans sağlayacağı sonucuna varılmıştır.

Tablo II'de *Öğrenci* ağının sonuçları sunulmuştur. *Öğrenci* ağının girdisi, hata anından önceki çerçeveleri içerir ve hata tahminleme için kullanılmaktadır. Tablodaki *1-Çerçeve* sütunu hata gerçekleşmeden 1 sn önceki çerçeveyi işaret eder.

1-Çerçeve sütunu için Tablo I ve Tablo II karşılaştırıldığında öğrenci ağı her kip için öğretmenden daha başarılıdır. Bunun ana sebepleri, bilgi genişletme konseptinden [28] ilham alınarak öğrenci ve öğretmen ağları bire bir aynı derinlikte seçilmiştir. Buna ek olarak öğrenci ağı hatayı görmemesine karşın öğretmen ağının hata için sahip olduğu tüm bilgiyi damıtmakta ve sadece bu bilgilerle yetinmeyip öğrendiği bilgileri gerçek veri etiketleriyle de kıyaslayarak doğrulamaktadır. Sonuç olarak öğrenci ağı; alışılmışın aksine [16], [17], [18] bu problem için öğretmen ağından daha küçük değil aynı boyutta seçilen, öğretmen ağının bilgisini hatayı tahmin etmek için damıtan ve gerçek veri etiketlerini de sürece dahil eden bir ağ olarak tasarlanmıştır. Bu sebeple *1-Çerçeve* için öğretmen ağından daha iyi bir başarıya sahiptir. *2-Çerçeve* ile *1-Çerçeve* karşılaştırıldığında, *1-Çerçeve* hata anına daha yakın olduğundan öğrenci ağının hata tahmin başarısı daha yüksektir.

IV. Sonuç

Sonuç olarak, hata tahminleme, otonom sistemlerin güvenilirliğini ve dayanıklılığını sağlamanın kritik bir adımıdır. Nesne etkileşimi izleme teknikleri kullanılarak, potansiyel hataları gerçek zamanlı olarak tespit etmek ve istenmeyen sonuçların ortaya çıkmasından önce önlem almak mümkündür. Önerilen sistem ile nesne etkileşim hataları 1 sn önceden %82.12, 2 sn önceden %79.79 başarı ile tahmin edilebilmektedir. Elde edilen sonuçlar, önerilen sistemin robot hatalarının önceden tahminlenmesinde kullanılabileceğini göstermektedir. Geliştirilen video dönüştürücü tabanlı mimarinin robot üzerindeki gerçek zamanlı testleri devam etmektedir.

Bilgilendirme

Bu çalışma, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından 119E436 Numaralı proje ile desteklenmiştir. Projeye verdiği destekten ötürü TÜBİTAK'a teşekkürlerimizi sunarız.

Kaynaklar

[1] M. Ersen, E. Oztop, and S. Sariel, "Cognition-enabled robot manipulation in human environments: Requirements, recent work, and open problems," *IEEE Robotics Automation Magazine*, vol. 24, no. 3, pp. 108–122, 2017.

[2] A. Inceoglu, E. E. Aksoy, A. C. Ak, and S. Sariel, "Fino-net: A deep multimodal sensor fusion framework for manipulation failure detection," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 6841–6847.

[3] A. Inceoglu, E. E. Aksoy, and S. Sariel, "Multimodal detection and classification of robot manipulation failures," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1396–1403, 2024.

[4] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder," *Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.

[5] D. Park, H. Kim, and C. C. Kemp, "Multimodal anomaly detection for assistive robots," *Autonomous Robots*, vol. 43, no. 3, pp. 611–629, 2019.

[6] S. Thoduka, J. Gall, and P. G. Plöger, "Using visual anomaly detection for task execution monitoring," in *International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4604–4610.

[7] P. Gohil, S. Thoduka, and P. G. Plöger, "Sensor fusion and multimodal learning for robotic grasp verification using neural networks," in *International Conference on Pattern Recognition*, 2022, pp. 5111–5117.

[8] D. Altan and S. Sariel, "What went wrong? identification of everyday object manipulation anomalies," *Intelligent Service Robotics*, vol. 14, no. 2, pp. 215–234, 2021.

[9] ——, "Clue-ai: A convolutional three-stream anomaly identification framework for robot manipulation," *Preprint arXiv:2203.08746*, 2022.

[10] A. C. Ak, E. E. Aksoy, and S. Sariel, "Learning failure prevention skills for safe robot manipulation," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 7994–8001, 2023.

[11] M. Hutchinson and V. Gadepally, "Video action understanding: A tutorial," *arXiv preprint arXiv:2010.06647*, 2020.

[12] Z. Zhong, M. Martin, M. Voit, J. Gall, and J. Beyerer, "A survey on deep learning techniques for action anticipation," *arXiv preprint arXiv:2309.17257*, 2023.

[13] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3118–3125.

[14] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1942–1950.

[15] A. Furnari and G. M. Farinella, "What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6252–6261.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from bert into simple neural networks," *arXiv preprint arXiv:1903.12136*, 2019.

[18] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

[19] G. Camporese, P. Coscia, A. Furnari, G. M. Farinella, and L. Ballan, "Knowledge distillation for action anticipation via label smoothing," in *International Conference on Pattern Recognition*, 2021, pp. 3312–3319.

[20] V. Tran, Y. Wang, Z. Zhang, and M. Hoai, "Knowledge distillation for human action anticipation," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2518–2522.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6836–6846.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[23] R. G. Lopes, S. Fenu, and T. Starner, "Data-free knowledge distillation for deep neural networks," *arXiv preprint arXiv:1710.07535*, 2017.

[24] X. Liu, X. Wang, and S. Matwin, "Improving the interpretability of deep neural networks with knowledge distillation," in *IEEE International Conference on Data Mining Workshops*, 2018, pp. 905–912.

[25] J. J. Gibson, "The perception of the visual world." 1950.

[26] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[27] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[28] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.