Gürültü İçeren Videolardan İnsan Hareketlerinin Çoklu Örnekle Öğrenme ile Tanınması
Recognizing Human Actions From Noisy Videos via Multiple Instance Learning

Fadime Sener, Nermin Samet, Pinar Duygulu
Bilgisayar Mühendisliği Bölümü
Bilkent Üniversitesi
Ankara, Türkiye
Email: fadime.sener,nermin.samet,duygulu@cs.bilkent.edu.tr

Nazli Ikizler-Cinbis
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
Ankara, Türkiye
Email: nazli@cs.hacettepe.edu.tr

***Özetçe* —Bu çalışmada videolardaki insan hareketlerinin tanınması, gürültünün tanıma performansına etkileri muhtemel bir çözüm yöntemiyle beraber incelenmiştir. Bilgisayarlı görü literatüründe mevcut olan veri kümeleri nispeten küçüktür ve etiketleme kaynağına bağlı olarak gürültü içerebilmektedirler. Oluşturulabilecek daha büyük veri kümelerinde ise gürültü artabileceği için, geleneksel öğrenme yöntemlerinin iyi performans sergilemeyecekleri bilinmektedir. Bu çalışmada veri kümelerindeki gürültünün artması durumunda yararlanılabilecek çoklu örnekle öğrenme tabanlı bir öğrenme yöntemi sunulmaktadır. Buna göre videolar, uzay-zaman ilgi noktaları ve görsel kelime kümeleri ile ifade edilmektedir. Daha sonra örnek tabanlı öğrenme ve çoklu örnekle öğrenme için destek vektör makineleri (DVM) kullanılarak sınıflandırıcılar oluşturulup, karşılaştırılmaktadır. Elde edilen sınıflandırma sonuçlarına göre, önerilen yöntemin gürültü içeren videolarda, örnek tabanlı öğrenme yöntemlerinden daha iyi performans sergilediği görülmektedir.**

Anahtar Kelimeler—insan hareketi tanıma; çoklu örnekle öğrenme; video anlama; veri gürültüsü

***Abstract*—In this work, we study the task of recognizing human actions from noisy videos and effects of noise to recognition performance and propose a possible solution. Datasets available in computer vision literature are relatively small and could include noise due to labeling source. For new and relatively big datasets, noise amount would possible increase and the performance of traditional instance based learning methods is likely to decrease. In this work, we propose a multiple instance learning-based solution in case of an increase in noise. For this purpose, each video is represented with spatio-temporal features, then bag-of-words method is applied. Then, using support vector machines (SVM), both instance-based learning and multiple instance learning classifiers are constructed and compared. The classification results show that multiple instance learning classifiers has better performance than instance based learning counterparts on noisy videos.**

Keywords—Human Action Recognition, Multiple Instance Learning, Video Understanding, Data Noise

978-1-4673-5563-6/13/$31.00 ©2013 IEEE

I. GİRİŞ

İnsan hareketleri tanıma, bilgisayarlı görünün önemli ve üzerinde çokça çalışılan alanlarından biridir. İnsan hareketi tanımanın amacı, sürmekte olan olayların otomatik olarak analizini sunmaktır. Videolardan insan hareketlerinin tanınmasına; gözetim sistemleri, hasta takip sistemleri, güvenlik, robotik vb. çeşitli insan ve makine etkileşimi içeren alanlarda ihtiyaç duyulmaktadır. Bununla beraber video tanıma içerisinde; arka plan değişimleri, kamera açısındaki değişikler, sınıf içi ve sınıflar arası değişkenlik vb. çeşitli zorlukları barındıran bir problemdir.

Bilgisayarlı görü literatüründe mevcut olan veri kümeleri el ile etiketlenmiş olmakla beraber veri toplama ve etiketleme zaman gerektiren bir işlem olduğundan, bu veri kümeleri nispeten küçük veri kümeleridir. Bununla beraber etiketleme sübjektif bir kaynağa, yani insana dayandığından, var olan veri kümelerinde yanlış etiketlerin bulunması da sıkça karşılaşılabilecek bir durumdur. Veri toplama işlemi, arama motorundan indirmek, video kaynaklarından indirmek vb. gibi çeşitli yöntemlerle otomatikleştirilmeye çalışılmıştır. Son zamanlarda, daha büyük veri kümelerinin otomatik olarak oluşturulması üzerine çalışmalar bulunmaktadır. Laptev ve diğerleri [1] veri kümesi oluştururken; filmler, filmlerin senaryolarını ve altyazılarını kullanarak hareketin bulunacağı kareleri videolardan otomatik olarak bulmakta ve veri kümelerini oluşturmaktadırlar (Hollywood2 veri kümesi). Buna göre oluşturdukları veri kümesinin içerdiği doğru video oranı yüksek olsa da; bu veri kümesi video, altyazı ve senaryo arasındaki eşleme bozukluğu gibi sebeplerden gürültü içermektedir. Olası bir çözüm, bu veri kümelerinin insanlar tarafından yeniden etiketlenmesidir, fakat bu çok zaman alan maliyetli bir işlemdir. Bu çalışmamızda bu veri kümesi ve ilerde otomatik olarak oluşturulabilecek veri kümeleri için yeni bir etiketlemenin sağlanmaması durumuyla başa çıkabilecek bir yöntem geliştirmeyi amaçladık.

Çoklu örnekle öğrenme, gürültü içeren problemler için uygun bir öğrenme yöntemidir. Geleneksel, örnek tabanlı öğrenmeden farklı olarak, bu öğrenmede teksel örnekler yerine, örneklerden oluşmuş torbalar kullanılır. Bu yöntemde torba etiketleri bilinirken, torba içindeki örneklerin etiketleri bilin-

memektedir. Torbanın etiketi belirlenirken; içerdiği örneklerden en az bir tanesi pozitif ise torba pozitif olarak etiketlenmekte; eğer bütün örnekler negatif ise torba negatif olarak etiketlenmektedir. Çalışmamızda, videolar birer örnek olarak kabul edilip, torbalar bu videolardan oluşturulmaktadır. Bu şekilde bütün torbaların en az bir pozitif örnek içereceği varsayılmaktadır.

Çalışmamızda videoları tanımlarken Laptev [2] tarafından tanımlanan ve birçok çalışmada başarılı sonuçlar vermiş olan uzay-zaman ilgi noktaları (STIP) öznitelikleri kullanılmıştır. Daha sonra bu öznitelikler başarılı performans gösterdiği bilinen görsel kelime kümeleri (bag-of-words) [3] olarak ifade edilmektedir ve bu gösterim üzerinden ilgili dağılımlar hesaplanmaktadır. Elde edilen bu dağılımlar üzerinden, destek vektör makinesi ile çeşitli çekirdekler ve çoklu örnekle öğrenme yöntemi kullanılarak her hareket sınıfı için ayrı sınıflandırıcılar oluşturulmaktadır. Elde edilen deney sonuçları, geleneksel gözetimli öğrenme yöntemleri ile karşılaştırıldığında, çoklu örnekle öğrenmenin veri kümesinin gürültü içermesi durumu ile daha kolay başa çıkabildiğini göstermektedir.

II. İLGİLİ ÇALIŞMALAR

Bu çalışmada gürültü içeren videolardan hareketleri tanıma amacıyla çoklu örnekle öğrenme kullanılmaktadır. Videolardan aktivite tanıma bilgisayarlı görünün çokça araştırılan bir konusudur. İnsan aktiviteleri karmaşıklıklarına, uzunluklarına göre; kol uzatmak gibi en küçük anlamlı insan hareketi olan vücut hareketleri, çok sayıda vücut hareketinden oluşan hareketler, bir insan ve insanın etkileşim içinde olduğu hareketler, bir insan ve nesnenin etkileşimde olduğu hareketler ve grupların hareketleri türlerinde olabilir. Bu çalışmada çok sayıda vücut hareketlerinden oluşan periyodik hareketler ele alınmıştır.

Hareket tanıma konusunda yapılmış çok sayıda çalışma mevcuttur. İlk çalışmalardan örnekler verecek olursak [4] ve diğerleri hareket analizindeki optik akışı temel alarak yöntemlerini sunmuşlardır. Bazı çalışmalar ise öznitelik takibine dayanmaktadır [5] [6]. Daha sonra yapılan çalışmalardan Nguyen ve diğerleri [7] karmaşık videolarda tanıma için hiyerarşik Gizli Markov Modele (Hidden Markov Model-HMM) dayalı istatiksel bir metot geliştirmişlerdir. Bunların yanında tanımlayıcı temelli yöntemler de mevcuttur [8].

Bu çalışmadakine benzer olarak, bazı çalışmalar videolardan insan hareketi tanıma problemi için videoları görsel kelime kümeleri olarak ifade etmişlerdir [9] [10]. Niebles ve diğerleri [11] bu çalışmadaki yönteme benzer şekilde videoları uzay-zaman ilgi noktaları ile tanımlayıp görsel kelime kümeleri ile sunmuşlar ve denetimsiz bir öğrenme yöntemi kullanarak hareketleri sınıflandırmışlardır. Laptev ve diğerleri de [1] uzay- zaman ilgi noktalarını görsel kelime kümeleri ile ifade edip çok kanallı lineer destek makineleri ile sınıflandırıcı oluşturup hareket tanıma yapmışlardır.

Çoklu örnekle öğrenme, son yıllarda veri madenciliği, sahne öğrenme, metin sınıflandırma gibi birçok uygulamada kullanılmıştır. Çoklu örnekle öğrenme ilk olarak Dietterich ve diğerleri [12] tarafından bir ilacın etkisini tahmin etme problemi için kullanılmıştır. Bu çalışmada ilaç tahmin etmedeki problemin temel zorluğu moleküllerin çok sayıda alternatif birleşme şekillerine sahip olmasıdır. Çoklu örnekle öğrenme kullanıldığında bu alternatif birleşme şekilleri birer örnek ve her molekül de bir torba olarak kabul edilip probleme çözüm getirilmiştir. Bu çalışmada kullandığımız yöntem ise Chen ve arkadaşları [13] tarafından geliştirilen Multiple-Instance Learning via Embedded Instance Selection (MILES) yöntemidir.

Şekil 1. HMDB51 veri kümesinden kullanılan beş hareket kategorisi için bazı örnek video kareleri.

III. VERİ KÜMESİ

Bu çalışmada [14] tarafından sunulan "HMDB51" veri kümesi kullanılmıştır. Bu veri kümesi çeşitli kaynaklardan; özellikle filmlerden, YouTube ve Google videoları gibi açık kaynaklardan elde edilmiş toplam 6849 adet video içermektedir. Toplanan bu veri kümesi en az 101 adet video içeren 51 hareket kategorisine ayrılmıştır. HMDB51 veri kümesindeki hareketler genel yüz hareketleri, nesne içeren yüz hareketleri, genel vücut hareketleri, nesne içeren vücut hareketleri ve etkileşimli vücut hareketleri olmak üzere 5 grupta toplanabilir. Bu veri kümesi için videodaki hareketin 51 sınıftan hangisine ait olduğu bilgisi etiketlenmiştir. Buna ek olarak bu veri kümesi için yazarlar tarafından 70/30 balansını ve çapraz doğrulamayı sağlayacak şekilde ayrılmış olan öğrenme ve test kümeleri bulunmaktadır.

Çalışmamızda bu veri kümesinden $brush\ hair$ (saç tarama), $dive$ (dalma), eat (yeme), $golf$ (golf) ve $ride\ horse$ (ata binme) olmak üzere 5 hareket kategorisi kullanılmaktadır. Performans değerlendirme aşamasında [14] tarafından sağlanan öğrenme ve test kümesi kullanılmıştır. Şekil 1'de kullanılan HMDB51 veri kümesinden alınan videolardan örnek kareler sunmaktayız.

IV. ÖNERİLEN YÖNTEM

Bu çalışmada hareketleri sınıflandırmak için önerilen yöntem şu aşamalardan oluşmaktadır. İlk adım olarak videolardan uzay-zaman ilgi noktaları öznitelikleri çıkarılmıştır. Daha sonra k-ortalamalar (k-means) kümeleme algoritması uygulanarak görsel kelimeler bulunup, dağılımlar hesaplanmıştır. Ardından örnek tabanlı öğrenme ve çoklu örnekle öğrenme için destek vektör makinesi kullanılarak, her hareket kategorisine ait sınıflandırıcılar oluşturulmuş ve performans karşılaştırması sunulmuştur. Son olarak da hareket kümelerinin gürültü içermesi durumunda, geleneksel öğrenme ve çoklu örnekle öğrenme karşılaştırılmıştır.

(t). kare

(t + a). kare

Şekil 2. Bir hareket videosuna ait iki farklı zamandaki uzay-zaman ilgi noktaları görülmektedir. Kamera değişiminden kaynaklı olarak t zamanında elde dilen ilgi noktaları gürültü içermektedir.

A. *Nitelik Çıkarma*

Uzay-zaman ilgi noktaları [2] önitelikleri son zamanlarda sıklıkla kullanılmakta olup performansları oldukça iyidir. Uzay zaman ilgi noktaları, Harris bulucusunun, 2 boyutlu uzaydan 3 boyutlu uzay-zamana genişletilmiş halidir. İlgi noktaları hem şekil hem de hareketteki değişime en büyük tepkiyi veren noktalardır. İlgi noktaları HOG (gradyen histogramları - Histogram of Oriented Gradients) ve HOF (optik akış histogramları - Histogram of Optical Flow) öznitelikleri ile tanımlanır. Şekil 2 'de veri kümesindeki bir videonun iki farklı zamanda çıkarılmış uzay-zaman ilgi noktalarını sunmaktayız. Uzay-zaman ilgi noktaları başarılı performansa sahip olsalar da kamera hareketi gibi durumlarda gürültü de oluşabilmektedirler.

Bu çalışmada her video görsel kelime kümeleri (bag of visual words) [3] ile ifade edilmektedir. Bu amaçla, yapılması gereken ilgi noktalarının bulunması ve bu ilgi noktalarının öznitelik vektörlerinin hesaplanmasıdır. Bu çalışmada uzay-zaman ilgi noktaları öznitelikleri kullanılmış ve her videoda değişik sayıda ilgi noktası olabileceği kaydı ile videolardan [162xN] boyutunda öznitelik matrisi elde edilmiştir. Uzay-zaman ilgi noktalarını elde etmek için HOG ve HOF betimleyicileri kullanılmaktadır. Buna göre HOG betimleyicileri 72 adet yerel bilgi içeren bileşen içermektedir, HOF betimleyicisi ise yerel hareket bilgisini içeren 90 bileşen içermektedir, toplamda her ilgi noktasına ait [162x1] boyutunda vektör elde edilmektedir. Öznitelikler k-ortalamalar kümeleme algoritması kullanılarak 1000 gruba kümelenmiştir. Bu kümelerin orta noktaları görsel kelimeleri oluşturmaktadır. Çıkarılan uzay-zaman ilgi noktaları, öznitelik vektörlerinin hesaplanmış kelime noktalarına Öklid uzaklığı hesaplanır ve her öznitelik vektörü, kendisine en küçük uzaklığa sahip görsel kelime ile eşleştirilir. Bu görsel kelime kümelerinin görülme sıklığı hesaplanarak kelime kümesi dağılımı (histogram) hesaplanır ve her video için 1x1000 boyutlu dağılımlar elde edilir.

B. *Çoklu Örnekle Öğrenme*

Çoklu örnekle öğrenme aşamasında, her video bir torba olarak kabul edilip ardından Chen ve diğerleri [13] tarafından geliştirilen Multiple Instance Learning via Embedded Instance Selection (MILES) algoritmasını kullanılmıştır. Bu yöntemde her torba veri kümesinde var olan örneklerle benzerliklerine göre yeni bir benzerlik uzayına taşınmaktadır. Torba $\mathbf{B}_i$ ve örnek c_l arasındaki benzerlik

$$s(c_l, \mathbf{B}_i) = \max_j \exp\left(-\frac{D(x_{ij}, c_l)}{\sigma}\right), \quad (1)$$

ile bulunmaktadır. $D(x_{ij}, c_l)$ örnek c_l ile her torba örneği x_{ij} arasındaki uzalıktır ve bu uzaklığın hesaplanmasında uygun olan herhangi bir uzaklık metriği kullanılabilir. Biz bu çalışmada, bu uzaklık ölçütü olarak Öklid uzaklığını kullanmaktayız.

Daha sonra, her bir torba, veri kümesindeki örnek noktalara benzerliğine göre şu şekilde gömülü bir temsil oluşturulur:

$$\mathbf{m}(B_i)) = [s(c_1, \mathbf{B}_i), \ldots, s(c_N, \mathbf{B}_i)]^T. \quad (2)$$

Yeni benzerlik uzayındaki bu gömülü temsil üzerinde, destek vektör makineleri ile sınıflandırıcılar öğrenilir. Her insan hareketi sınıfı için ayrı ayrı öğrenilen sınıflandırıcılar, deney aşamasında, teker teker deney videoları üzerinde uygulanır ve her deney videosunun etiketi olarak en yüksek çıktıyı veren sınıflandırıcının sınıfı atanır.

V. DENEYLER

Çalışmamızın sonuçları ortalama kesinlik (Average Precision) kriterine göre değerlendirilmiştir. Kesinlik, kullanılan yöntemin tahmin ettiği doğru pozitiflerin yüzdesidir.

Çalışmamızda öncelikle gürültü içermeyen veri kümesi, örnek tabanlı geleneksel öğrenme ve çoklu örnekle öğrenme kullanılarak test edilmiştir. Örnek tabanlı öğrenme için, 5 kategori için destek vektör makineleri (DVM) kullanılarak ayrı ayrı sınıflandırıcı modelleri öğrenilmiştir. Destek vektör makinelerinin farklı problemlerde farklı çekirdeklerle (kernel) kullanılmasının performansı etkilediği bilinmektedir. Bu amaçla bu çalışmada, farklı çekirdek kullanımları denenmiş, RBF, Polynomial ve Hellinger çekirdek fonksiyonları kullanarak sonuçlar elde edilmiştir. En iyi sonuçlar Hellinger çekirdek ile elde edilmiştir. İki h ve h' dağılımı için Hellinger çekirdek Formül 3 'deki gibi ifade edilmektedir.

$$k(h, h') = \sum_i \sqrt{h(i)h'(i)} \quad (3)$$

Tablo I'de sonuçlarımızı sunmaktayız. Hareket sınıflandırma doğruluk oranlarının özellikle "yeme" hareketi için düşük olduğunu gözlemliyoruz; bu durumu hareketin karmaşıklığına ve varyasyonun çokluğuna bağlayabiliriz.

Sonraki adım olarak çoklu örnekle öğrenme yöntemini test etmekteyiz. Buna göre çoklu örnekle öğrenme için önemli bir parametre olan torba boyunu belirlemek ve en iyi torba boyunu elde etmek için, her torba k = 3, 6 ve 9 video içermek

Tablo I. HMDB51 VERI KÜMESI ÜZERINDE HELLINGER ÇEKIRDEK KULLANAN DESTEK VEKTÖR MAKINELERI ILE SINIFLANDIRMA DOĞRULUK SONUÇLARI.

yöntem	Saç Tarama	Dalma	Yeme	Golf	Ata Binme	ORTALAMA
DVM (Hellinger)	87.78	87.52	52.95	67.59	73.44	73.86

Tablo II. ÇOKLU ÖRNEKLE ÖĞRENME TORBA BOYU SEÇIMI

Kategori	k=3	k=6	k=9
Saç Tarama	85.73	75.10	73.79
Dalma	89.62	83.79	82.42
Yeme	57.58	45.93	43.56
Golf	74.04	70.38	66.24
Ata Binme	85.81	73.08	72.55
ORTALAMA	**78.56**	69.66	67.71

kaydıyla test edilmiştir, ve elde edilen deney sonuçları Tablo II'de sunulmaktadır. Torba boyu olarak 3 seçilmesi durumunda en yüksek performansa sahip olduğumuzu söyleyebiliriz. Buna ek olarak çoklu örnekle öğrenmenin veri kümesi üzerinde gösterdiği performans, destek vektör makinesi ve Hellinger çekirdek kullanılmasının performansından daha iyidir. Çoklu örnekle öğrenme 5 hareket için performansı artırmakla beraber "yeme" hareketinin performansı hala düşüktür. Veri kümesindeki videoların yeterince iyi olmaması durumunda örnek tabanlı sınıflandırmanın olumsuz etkilenebileceği bilinmektedir, fakat bu zayıf video örnekleri çoklu örnekle öğrenme kullanıldığında sınıflandırmaya daha az negatif etki etmektedir. "eat" hareketi için performansın düşük olmasını karışık bir hareket olmasına ve kullandığımız öznitelikler tarafından iyi bir şekilde ifade edilmemesinden kaynaklandığını düşünmekteyiz.

Veri kümesine %10 gürültü eklediğimizde elde edilen sonuçlar Tablo III'de sunulmaktadır. Gürültü veri kümesi, HMDB51 veri kümesinin bu çalışmada seçilen 5 kategori haricindeki video kategorilerinden rastgele seçilmiştir. Geleneksel öğrenme gürültü eklenmediği durumda bile çoklu örnekle öğrenmeden kötü performans sergilemekteyken, gürültü eklendiği durumda performansı çok düşmüştür. Çoklu örnekle öğrenme beklenildiği gibi gürültüye karşı dayanıklılık gösterip, gözle görünür bir üstünlük sağlamıştır.

VI. SONUÇ

Bu çalışmada veri kümelerinin gürültü içermesi durumunda, örnek tabanlı yöntemler yerine, çoklu örnekle öğrenme kullanılmasının performansı arttırdığını göstermekteyiz. Elde edilen deneysel sonuçlar, artan veri kümesi ihtiyacı ile oluşacak etiketleme problemlerinde, etiketlemenin olmadığı otomatik veri kümesi oluşturulması durumlarında çoklu örnekle öğrenme kullanımının yararlarını göstermektedir. Bazı hareket sınıflarında yeterince iyi performans elde edememekle beraber bunun videonun optimal betimleyicilerle kodlanmamasından kaynaklandığını ve daha iyi tanımlayıcılar ile zenginleştirilebileceğini, ek olarak yöntemin veri kümesi olarak otomatik oluşturulmuş veri kümeleri üzerine uygulanabileceğini düşünmekteyiz.

Tablo III. %10 GÜRÜLTÜ EKLENMIŞ HMDB51 VERI KÜMESINDE ÖRNEK TABANLI ÖĞRENME VE ÇOKLU ÖRNEKLE ÖĞRENME SONUÇLARI

Kategori	DVM (Hellinger)	MILES
Saç Tarama	55.33	85.16
Dalma	74.13	85.11
Yeme	32.99	56.61
Golf	38.89	74.40
Ata Binme	46.81	84.11
ORTALAMA	49.63	**77.08**

KAYNAKÇA

[1] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies." in *CVPR*. IEEE Computer Society, 2008.

[2] I. Laptev, "On space-time interest points." *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[3] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *Proceedings of the International Conference on Computer Vision*, 2005.

[4] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques." *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[5] S. M. Smith and M. Brady, "Asset-2: Real-time motion segmentation and shape tracking." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 814–820, 1995.

[6] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," 1998.

[7] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden markov models." in *CVPR (2)*. IEEE Computer Society, 2005, pp. 955–960.

[8] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos." in *CVPR*. IEEE, 2009, pp. 2012–2019.

[9] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach." in *ICPR (3)*, 2004, pp. 32–36.

[10] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, October 2005.

[11] J. C. Niebles, H. Wang, and F.-F. Li, "Unsupervised learning of human action categories using spatial-temporal words." in *BMVC*. British Machine Vision Association, 2006, pp. 1249–1258.

[12] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles." *Artif. Intell.*, vol. 89, no. 1-2, pp. 31–71, 1997.

[13] Y. Chen, J. Bi, and J. Z. Wang, "Miles: Multiple-instance learning via embedded instance selection." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, 2006.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.