META-SINIF TABANLI GETİRME İLE VERİYE DAYALI İMGE ALTYAZILAMA
DATA-DRIVEN IMAGE CAPTIONING WITH META-CLASS BASED RETRIEVAL

Mert Kılıçkaya[1], Erkut Erdem[1], Aykut Erdem[1], Nazlı İkizler Cinbiş[1], Ruket Çakıcı[2]

[1]Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
{mert.kilickaya,erkut,aykut,nazli}@cs.hacettepe.edu.tr

[2]Bilgisayar Mühendisliği Bölümü
Orta Doğu Teknik Üniversitesi
ruken@ceng.metu.edu.tr

ÖZETÇE

Otomatik imge altyazılama, bir imgenin açıklamasını yaratma işlemi, bilgisayarlı görü ve doğal dil işleme topluluklarının ilgisini daha yeni çeken çok zorlu bir problemdir. Bu çalışmada, verilen bir imge için; imge-altyazı ikilileri içeren geniş bir veri kümesinden ona görsel olarak en benzer imgeyi bulan ve onun altyazısını girdi imgesinin açıklaması olarak aktaran veriye dayalı özgün bir imge altyazılama stratejisi önerilmiştir. Özgünlüğümüz, getirme işlemi için girdi görüntüsünün anlamsal içeriğini daha iyi yakalamak için meta-sınıg gösterimi olarak adlandırılan yeni önerilmiş yüksek düzey bir global imge gösterimi kullanılmasında yatmaktadır. Deneylerimiz meta-sınıf güdümlü yaklaşımımızın dayanak Im2Text modeline kıyasla daha doğru açıklamalar ürettiğini göstermektedir.

Anahtar kelimeler — görüntüden-metin, imge altyazılama

ABSTRACT

Automatic image captioning, the process of producing a description for an image, is a very challenging problem which has only recently received interest from the computer vision and natural language processing communities. In this study, we present a novel data-driven image captioning strategy which, for a given image, finds the most visually similar image in a large dataset of image-caption pairs and transfers its caption as the description of the input image. Our novelty lies in employing a recently proposed high-level global image representation, named the meta-class descriptor, to better capture the semantic content of the input image for use in the retrieval process. Our experiments show that, as compared to the baseline Im2Text model, our meta-class guided approach produces more accurate descriptions.

Keywords — image-to-text, image captioning

1. GİRİŞ

Miktarı her geçen gün artan bir hızla çoğalan İnternet'teki görsel veriler genellikle onlarla ilişkilendirilmiş açıklamalar ve altyazılar gibi metinsel veriler ile birlikte bulunmaktadır. Bu metin ve alt yazılarının kimilerinin imgenin görsel içeriğini açıklamaya yönelik olması, iki veri çeşidi arasında çok biçimli (multi-modal) mantıksal bir ilişkinin olduğunu göstermektedir. Bu yararlı birliktelik daha önceleri nesne tespiti ve imge etiketleme [1,2] gibi bilindik problemlerin çözümünde başarılı bir şekilde kullanılmıştır. Son yıllarda ortaya konan kimi yeni çalışmalar imge, etiketler ve kelimeler arasında kurulan başarılı bağları bir adım öteye taşıyarak verilen bir imgeyi açıklayan insan benzeri doğal bir açıklamanın otomatik olarak çıkarılması problemi üzerine gitmişlerdir [3-9]. Bu yöntemlerin tamamı, dil ve görme arasındaki bağdan yola çıkarak bilgisayarlı görü ve doğal dil işleme yöntemlerini bütünleşik bir yapıda bir araya getirmekte ve bu sayede çok biçimli (görsel ve metinsel) bilgiden tam anlamıyla faydalanmayı amaçlamaktadır.

Yukarıda bahsedilen *görüntüden-metin (image-to-text)* olarak adlandırabileceğimiz araştırma konusu beraberinde açıklayıcı metin içeren pek çok farklı imge veri kümelerinin ortaya çıkmasını da sağlamıştır. Bu tür çok biçimli veri kümelerinin ilk örneklerin olan ve ilk olarak Farhadi vd. [3] tarafından kullanılan *Pascal Sentences* [10], *Pascal Visual Object Class Challenge 2008* [11] veri kümesindeki bazı sınıflara ait imgeler için deneklere sorulup oluşturulan beşer farklı metinsel açıklama içermektedir. Çalışmamızda da kullandığımız bu denektaş veri kümesine ait bazı örnek görüntüler ve bunlarla ilişkili alt yazı açıklamaları Şekil 1'de verilmiştir. Bir diğer görece çok daha büyük ve daha doğal olan veri seti de [5] çalışması kapsamında oluşturulan, Flickr sitesinden 1 milyon imge ve onlara ait altyazıların toplanmasıyla oluşturulmuş *SBU Captioned Photo Set* adlı veri kümesidir.

Bu çalışmada Ordonez vd. tarafından önerilen *Im2Text* adlı yöntem [5], yakın tarihte önerilmiş olan *meta-sınıf (meta-class)* [12] öznitelikleri üzerinden yeniden ele alınmıştır. Im2Text çalışması, görüntülerden metinsel açıklama yaratmayı kısaca bir bilgi çekme (information retrieval) problemi olarak ele almaktadır. Bu amaçla ilk olarak tasviri oluşturulacak sorgu imgesi için veri kümesinde ona benzer olan imgeler alt düzey global gösterimler üzerinden getirilmekte ve ardından bu imgeler üzerinden çıkarılan yerel üst düzey bilgilere göre en yakın imgenin açıklaması ilgili sorgu imgesinin açıklaması olarak transfer edilmektedir. Bu çalışmada amacımız, çıkarılacak açıklamanın başarısını doğrudan etkileyen ilk getirme işleminin direkt üst düzey bilgiler kullanan bir global gösterim ile daha iyileştirilmesidir.

Bildirinin geri kalanı şu şekilde düzenlenmiştir: Bölüm 2'de öncelikle imge gösteriminde kullanılan meta-sınıf gösterimi [12] açıklanmakta ve bu gösterime dayalı olarak tasarladığımız altyazı transferi yönteminin detayları verilmektedir. Daha sonra Bölüm 3'te geliştirdiğimiz yöntemin başarımı ölçülerek önceki yöntemlerle kıyaslanmakta, ve ardından Bölüm 4'te deneyler sonucunda elde edilen sonuçlara yer verilmektedir.

2. YÖNTEM

Önerdiğimiz yaklaşım, imge-cümle ilişkisine dair insanların bir imgeye açıklama getirirken imgenin konusunu oluşturan nesne veya nesnelere önem verdikleri gözlemine dayanmaktadır. Kısaca, içerdikleri nesnelere bağlı olarak bir imgenin daha üst düzeyde bir temsilinin elde edilmesi ve

978-1-4799-4874-1/14/$31.00 ©2014 IEEE

- An old man with sheep in the background.
- An old shepherd in the mountains
- A picture of a man with some sheep in the background.
- Older man wearing beret with mountains in background.
- Old man with sheep

- A man holds a ball in a puppies mouth.
- A puppy bites a ball.
- A small puppy being fed a chocolate treat.
- A tan puppy with a hand holding something in his mouth.
- Someone is putting something in the white dog's mouth.

- A bicycle racer on a road in a rural area
- A man in green and yellow lira riding a bike through the countryside.
- A man on a bicycle with a racing suit.
- Cyclist pedaling down country road.
- The cyclist is speeding along on a country road in his yellow and green suit.

- A city bus driving past a building.
- A red bus picks up new passengers
- A red trolley bus passing by on the opposite side of a city street.
- Red bus in front of building.
- The bus makes the rounds in the quaint town.

Şekil 1. Pascal Sentences veri kümesinden bazı örnek imgeler ve bu imgelere ait cümlesel tanımlamalar.

bunun üzerinden imgenin açıklamasının otomatik olarak çıkarılması önerilmiştir.

Bu üst düzey temsil, görüntü içeriğini anlamsal olarak ifade etmeyi gerektirmektir ve bundan dolayı son yıllarda popülerlik kazanan nitelik bazlı [13] gösterimlerle de ilişkilendirilebilir. Verilen bir imgenin imgede mevcut nesnelere dayalı bu tarz bir gösteriminin çıkarımı için literatürde yakın tarihli bazı çalışmalar yapılmıştır [12, 14, 15]. Çalışmamızda, bu temsil için mevcut modeller arasından sınıflandırmadaki başarısından hareketle Bergamo ve Torresani tarafından önerilmiş *meta-sınıf* modeli [12] kullanılmıştır.

Meta-sınıf modeli, geleneksel nitelik bazlı yaklaşımların kullandığı elle seçilmiş nesne sınıfları veya görsel özelliklere bağlı sınıflandırıcılar yerine düşük düzey görsel özniteliklere bağlı olarak öğrenilmiş doğrusal olmayan sınıflandırıcıların çıktılarını verilen bir imgeyi ifade etmek için kullanmaktadır. Bu sınıflandırıcılar, gerçek dünyada mevcut olması gerekli olmayan, benzer özellikleri taşıyan nesne sınıflarını (örneğin otobüs ve arabalar tarafından paylaşılan tekerlek onları taşıt kategorisine dahil eder) kendi içinde kapsayan soyut sınıfları yakalayacak şekilde tanımlanmıştır ve bundan dolayı sınıflar ötesi anlamsal özellikleri yakalamaktadırlar. İlgili çalışmada bu sınıflandırıcıların öğrenimi, *ImageNet* [16]'in bir altkümesi üzerinden uzlamsal piramit histogramlarına dayalı olarak GIST [17], HOG [17], SSIM [19] ve SIFT [20] öznitelikleri kullanılarak gerçekleştirilmiştir.

Verilen bir imgenin otomatik açıklama çıkarımı için [5]'de olduğu gibi örnek tabanlı, veri güdümlü bir yol izlenmektedir. Ayrıntılı olarak açıklamak gerekirse, bu girdi imgesi için öncelikle meta-sınıf tabanlı bir gösterimin elde edilir:

$$\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_C(\mathbf{x})]^T \quad (1)$$

Burada $\mathbf{x}$ girdi imgesini betimleyen öznitelik vektörünü, h_i (i = 1,...,C)'ler ise temsilde kullanılan tekil meta-sınıf sınıflandırıcılarının çıktılarını ifade etmektedir.

Bu işlemin ardından açıklama çıkarımı için; elde edilen bu temsile dayalı olarak imge-cümleler ikilileri içeren bir veritabanı içinde bir yakınlık sıralaması yapılmaktadır ve getirilen en yakın görüntünün girdi görüntüsüne anlamsal içerik olarak en benzer görüntü olduğu varsayımı üzerinden onun altyazısı ilgili girdi imgesinin metinsel açıklaması olarak transfer edilmektedir.

3. DENEYLER

3.1 Veri Kümesi

Önerilen yöntemin başarı değerlendirmesi amacı ile giriş bölümünde değindiğimiz Pascal Sentences [10] veri kümesi kullanılmıştır. Bu veri kümesinde, 1000 tane imge ve her bir imgeye ait 5 farklı kişi tarafından eklenmiş imge açıklaması, toplamda 5000 imge açıklaması bulunmaktadır. Bu veri seti, orjinal Pascal nesne tespit ve sınıflandırma veri kümesinin bir alt kümesi olarak oluşturulmuş olup, 20 farklı sınıftaki nesneye ait 50 imgenin rastgele seçilmesi ile derlenmiştir. İmge açıklamaları, gerek dilbilgisi yapısı, gerekse içerik olarak farklılık gösterebilmektedir. Şekil 1'de bu veri kümesinden örnek imgeler ve bu imgelere ait açıklamalar verilmiştir. Görüldüğü gibi, her kişinin imgeye bakış açısı ve tanımlama şekli farklı olabilmektedir. Bu nedenle, bu bildiri kapsamında ele alınmakta olan imgelerin otomatik olarak tanımlanması ve açıklayıcılarının oluşturulması oldukça zorlu bir problemdir.

3.1. İmge Getirme

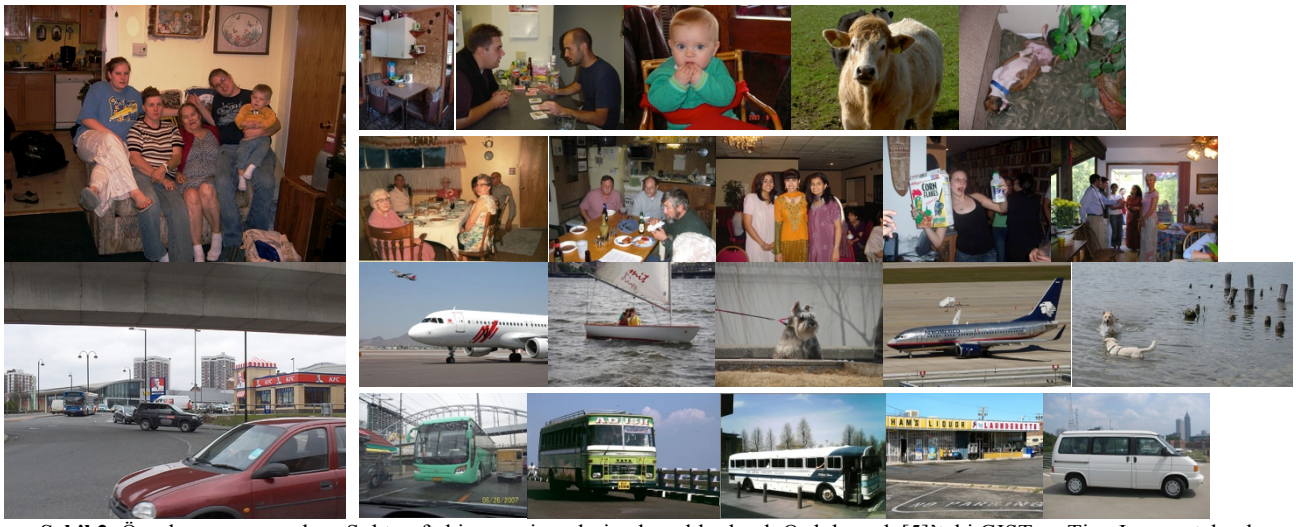
Yapılan ilk deneylerde aynı nesne sınıfına ait imgelerin geri getirilmesinin başarımı, [5]'de dayanak model olarak kullanılan GIST [17] + Tiny Images (TI) [21] temsilinin performansı ile kıyaslamalı bir şekilde ölçülmüştür. Birini dışarıda bırak çapraz geçerlemesi (leave-one-out cross validation) ile 20 farklı sınıf üzerinde nesne geri getirim kesinlikleri değerlendirilmiştir. Geri getirme başarımı ortalama kesinlik ölçümleriyle yapılarak sınıf bazında bu yöntemle hangi sınıfların daha başarılı geri getirildiği, hangilerinde karmaşanın arttığı gözlemlenmeye çalışılmıştır.

Şekil 2'de imge geri getirme deneylerinde kullanılan iki örnek imge için veri tabanında bulunan onlara en yakın 5'er imge listelenmektedir. Şeklin sol tarafında sorgu olarak verilen imgeler için, anlamsal analiz tabanlı meta-sınıf özniteliklerine dayalı geri getirmenin, ilintili imgeleri getirmede çok daha başarılı olduğu görülmektedir.

Tablo 1'de ilgili niceliksel sonuçlar, sınıf bazında hesaplanan ortalama geri getirim-kesinlik eğrisi altındaki alan (area under the precision-recall curve) (AUCPR) üzerinden sunulmuştur. Bu sonuçlar incelendiğinde, meta-sınıf üst düzey öznitelikleri yoluyla resimlerin anlamsal gösterimi ile gerçekleştirilen geri getirimin, bu işlemin genel sahne üzerinden çıkarılan alt düzey özniteliklere dayalı olarak yapılmasına oranla daha başarılı olduğu görülmüştür. Bu veri kümesinde çıkan genel olarak düşük değerler, veri kümesindeki imgelerin zorluğunu göstermektedir.

Tablo 1. Pascal Sentences veri kümesinde nesne bazında AUCPR değerleri.

	Aero-plane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	
Im2Text (GIST+TI)	0.11	0.05	0.09	0.09	0.05	0.04	0.04	0.04	0.04	0.04	
Önerilen (Meta-sınıf)	**0.27**	**0.07**	**0.06**	**0.14**	**0.08**	**0.18**	**0.09**	**0.10**	**0.07**	**0.10**	
	Diningtable	Dog	Horse	Motor bike	Person	Potted plant	Sheep	Sofa	Train	Tv monitor	**Ortalama**
Im2Text (GIST+TI)	0.05	**0.06**	0.05	0.05	0.04	0.04	0.06	0.05	0.04	0.04	0.054
Önerilen (Meta-sınıf)	**0.13**	**0.06**	**0.10**	**0.10**	**0.07**	**0.05**	**0.13**	**0.09**	**0.10**	**0.16**	**0.108**



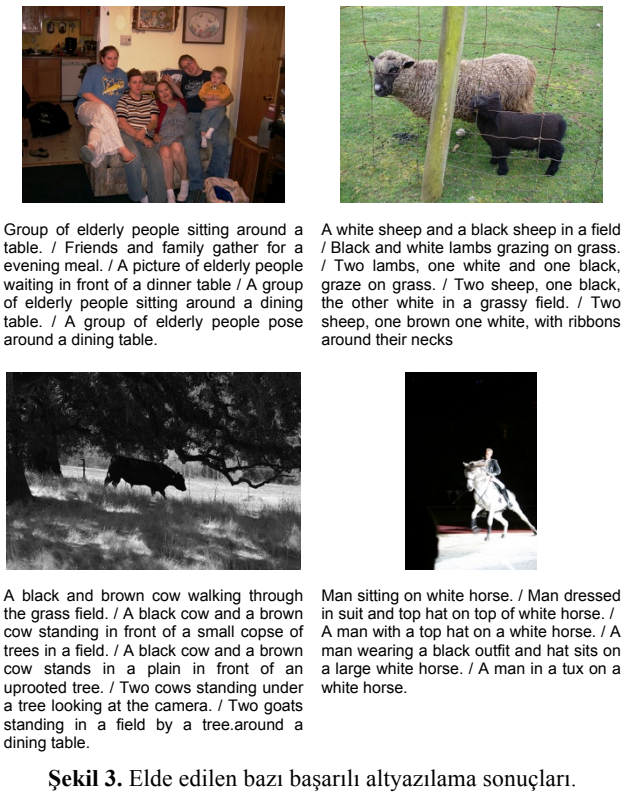
Şekil 2. Örnek sorgu sonuçları. Sol taraftaki sorgu imgelerine karşılık olarak Ordolez vd. [5]'teki GIST ve Tiny Images tabanlı yöntemin getirdiği imgeler sağ üst satırda; meta-sınıf tabanlı önerilen yöntemin getirdiği imgeler sağ alt satırda gösterilmektedir.

3.2. İmge Altyazılama

Bir sonraki aşamada, Im2Text [5] çalışmasına benzer bir yaklaşımla, verilen bir sorgu imgesinin görsel gösterimine en yakın geri getirilen imgeye ait alt yazı, sorgu imgesine transfer edilmiş ve başarım sorgu imgesinin esas alt yazısı ile olan BLEU skoru [23] ölçülerek yapılmıştır. Aslen makine çevrimi ve otomatik özet çıkarma sistemlerinin başarım analizinde kullanılan BLEU skoru, görüntüden metin oluşturma çalışmalarının başarımlarının niceliksel ölçümünde de sıkça başvurulan metriklerin başında gelmektedir [4-9] ve yaratılan açıklama ile referans açıklama arasındaki n-gram keskinliğine dayanmaktadır. Deneylerimizde Ordolez ve diğerlerinin [5] çalışmasının kullandığı GIST ve Tiny Images özniteliklerini temel alan gösterimi, bu çalışmada önerilen metaclass tabanlı geri getirme yöntemi ile kıyaslanmıştır. Elde edilen unigram BLEU skorları (BLEU-1), Tablo 2'de görülmektedir. Metaclass tabanlı anlamsal imge getiriminin her aşamada, [5]'da önerilen GIST ve Tiny Images (TI) tabanlı yaklaşımdan daha başarılı çalıştığı gözlemlenmektedir. Önerilen yaklaşım kullanılarak altyazılanan bazı başarılı ve başarısız sonuçlar Şekil 3 ve Şekil 4'de verilmiştir.

Tablo 2. İmge altyazılama BLEU-1 skorları

Yöntem	Minimum	Maksimum	Ortalama
Şans	0.0033	0.2589	0.0917
Im2Text (GIST+TI)	0.0047	0.2871	0.1060
Önerilen (Meta-sınıf)	**0.0067**	**0.3231**	**0.1181**



Şekil 3. Elde edilen bazı başarılı altyazılama sonuçları.

White Cunard cruise ship on a cloudy day. / The liner prepares to set out in stormy weather. / A large yacht boat called "Cunard" is docked. / A large ship is docked on a cloudy day. / A large ship is docked on a cloudy day. / A cruise ship in harbor with rain clouds overhead.

A closeup headshot of two women with one sticking out her tongue. / A woman in a gray shirt smiles for the camera while the woman behind her makes a face. / A woman with a surprised look on her face is holding a smiling woman from behind. / Two women make faces at the camera. / Two women posing for the camera with the women behind pulling a face

Şekil 4. İki başarısız altyazılama sonucu.

4. SONUÇLAR

Bu çalışmada imgelerin içeriğini açıklamaya yönelik otomatik altyazı üretimi için veri tabanlı yeni bir yöntem önerilmiştir. Bu yöntemde, veri kümesinden sorgulama ve geri getirme aşamasında, alt düzey imge özniteliklerinin yanısıra, anlamsal boyutta üst düzey bilgileri içeren meta-sınıf özniteliklerinin kullanılması önerilmektedir. Yapılan deneylerde, meta-sınıf açıklayıcıların, sadece alt düzey imge özniteliklerine dayalı getirim ve altyazılamaya oranla daha başarılı çalıştığı gözlenmektedir.

Yaklaşımız, veriye dayalı olduğu için başarısının kullanılan veri tabanının büyüklüğü ile ilişkili olduğu söylenebilir. Bu bakımdan deneylerde kullanılan veri kümesinin boyutunun görece küçük olmasından dolayı meta-sınıf gösteriminin, girdi görüntüsüyle aynı sınıftan imgeleri getirirken bunların yanında benzer sınıflara ait imgeleri de getirdiği (örneğin köpek sınıfı için onun meta-sınıfı olan hayvanlar kategorisine ait inek, kuş vb. sınıfları) gözlemlenmiştir. Bu durum, sorgu yapılan ve geri getirilen imgeler arasında, BLEU skoru bazında düşük benzerliklere sebep olmuştur. Benzer şekilde, örneğin oturma odası sınıfına ait bir imgede bulunan insanlar, insan sınıfına ait imgelerin geri getirilmesine sebep olmuş, mantıksal olarak olumlu sonuçları olsa da sayısal olarak alınan ölçümleri düşürmüştür. Bu bağlamda, gelecekte çalışmamızda SBU Captioned Photo Set [5] gibi çok daha geniş bir veri kümesinin kullanılması ve bu tarz problemlerin ortadan kaldırılması planlanmaktadır.

Ayrıca, bu durumlar için, ileriki çalışmalarda, farklı düzeydeki özniteliklerin sisteme entegrasyonu planlanmaktadır. Bu çalışmanın sonraki aşamalarında, insanların imgelere getirdiği açıklamalarla görüntüde baktıkları yerlerin bağını kuran ve imgenin her bölgesini kapsayan bir gösterim yerine, göreli olarak önemli yerlerin öncelikli olarak altyazı oluşturmada kullanılması planlanmaktadır. Yun vd. [23], insanların imgelere getirdiği açıklamalarla görüntüde baktıkları yerlerin bağını incelenmiş ve bu ikisi arasında önemli bir bağın olduğu göstermişlerdir.

5. TEŞEKKÜR

Bu çalışma kısmen TÜBİTAK-COST 113E116 nolu proje tarafından desteklenmiştir.

6. KAYNAKÇA

[1] Y. Mori, H. Takahashi, R. Oka, Image-to-word transformation based on dividing and vector quantizing images with words, In *WMISR,* 1999.

[2] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, In *ECCV,* 2002.

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, Every picture tells a story: Generating sentences from images, In *ECCV*, 2010.

[4] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, Baby talk: Understanding and generating simple image descriptions, In *CVPR*, 2011.

[5] V. Ordonez, G. Kulkarni, and T. L. Berg, Im2Text: Describing Images Using 1 Million Captioned Photographs, In *NIPS*, 2011.

[6] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, Composing simple image descriptions using web-scale n-grams, In *CoNLL*, 2011.

[7] Y. Yang, C. Teo, H. Daume III, and Y. Aloimonos, Corpus-guided sentence generation of natural images, In *EMNLP*, 2011.

[8] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daume III, Midge: Generating image descriptions from computer vision detections, In *EACL*, 2012.

[9] A. Gupta, Y. Verma, and C. Jawahar, Choosing linguistics over vision to describe images, In *AAAI*, 2012.

[10] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting Image Annotations Using Amazon's Mechanical Turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk,* 2010.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challange 2008 (VOC 2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop

[12] A. Bergamo, L. Torresani, Meta-Class Features for Large-Scale Object Categorization on a Budget, In *CVPR*, 2012.

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[14] L. Li, H. Su, E. Xing, and L. Fei-Fei, Object Bank: A high-level image representation for scene classification & semantic feature sparsification, In *NIPS*, 2010.

[15] L. Torresani, M. Szummer, and A. Fitzgibbon, Efficient object category recognition using classemes, In *ECCV*, 2010.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, In *CVPR*, 2009.

[17] A. Oliva and A. Torralba, Building the gist of a scene: The role of global image features in recognition, *Visual Perception, Progress in Brain Research*, 155, 2006.

[18] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, In *CVPR*, 2005.

[19] E. Shechtman and M. Irani. Matching local self-similarities across images and videos, In *CVPR*, 2007.

[20] D. Lowe. Distinctive image features from scale-invariant keypoints, *IJCV*, 60(2):91–110, 2004.

[21] A.Torralba, R.Fergus, and W.Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition, *IEEE T-PAMI*, 30, 2008.

[22] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, In *ACL*, 2002.

[23] K. Yun, Y. Peng, G. Zelinsky, D. Samaras, T. L Berg, Studying Relationships Between Human Gaze, Description, and Computer Vision, In *CVPR*, 2013.