Image Search

Sung-eui Yoon
KAIST

September 26, 2018

Chapter 9

Unsupersized and Weak-supervised Learning

While we have seen significant advances on various visual learning tasks, most of them are based on supervised learning, requiring a high number of annotated data. To achieve scalability for various tasks, even for internet-scale problems that have billions of images without labels and even with noisy and incorrect information, it is important to learn something out of non-supervised way. In this chapter, we discuss various unsupervised learning techniques.

9.1 Predicting Spatial Configuration

Utilizing many unlabelled images in an unsupervised manner has been not demonstrated to extract useful information. On the other hand, utilizing context has been successful in the text domain. For example, given a few words before and after, we can predict a word based on its context. This effectively transforms the unsupervised problem, learning a similarity function between two words, into a self-supervised problem , learning something on a word given its context.

Based on this success on utilizing the context in the text domain, we would like to design a similar process for images. In this way, we aim to predict a relative location of a patch from another patch, both of which are randomly extracted from an image (Fig. 9.1). Out of this process, we aim to learn useful semantic features that can be also useful for other tasks.

One can utilize an existing CNN architecture, but we should give care on avoiding trivial or hacky solutions for the architecture. For example, the deep learning can easily predict the relative positions based on shared edges and textures. To avoid this problem, patches for training are generated with some gaps and random positional jitterring between two extracted patches. Nonetheless, chromatic aberration, different frequency of lights behave differently with camera lenses, can be also a powerful cue on learning the relative position. One can adjust such chromatic aberration or use only a single color channel.

We can then use such trained networks as an unsupervised pre-training on some of computer vision tasks. The learned features from predicting the relative patch locations are used as an pre-training for the object detection problem, and shown to work better than randomly initialized networks. Nonetheless, there are still accuracy gaps between the unsupervised one and the pre-training network with labelled data.

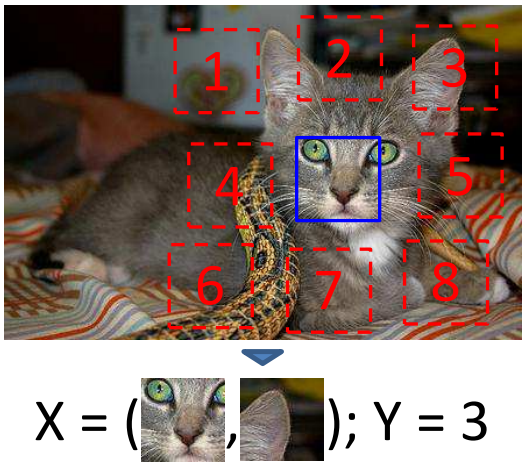


Figure 9.1: We consider a problem of predicting a relative position of a patch given another patch, both of which are randomly extraccted from an image. This image is excerpted from [DGE15]

9.2 Context Encoder: Feature Learning by Inpainting

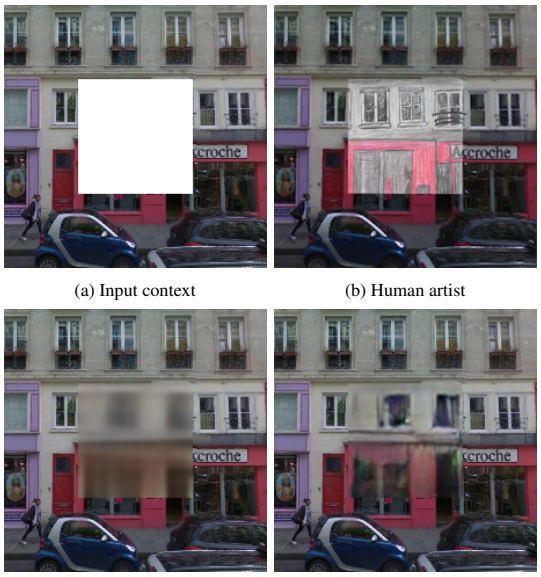
Inpainting is used for removing unwanted objects in images. Many different techniques such as nearest patch based approaches have been developed for inpainting. In this section, we discuss to utilize the inpainting process for learning features in an unsupervised manner.

To perform inpainting well, we need to understand various semantics and structures of images and predict on missing areas only from partially observed areas (Fig. 9.2).

One can imagine to use deep learning to perform such inpaining process. Context encoder is introduced to perform such process [PKD+16]. Furthermore, the context encoder can learn compact features in an unsupervised manner and thus can be used for many other high-level tasks as pre-training or unsupervised steps.

The context encoder shares a similar architecture to a common autoencoder consisting of encoder and decoder (Fig. 9.3). We can use the regular L2 loss between the predicted image region and its ground truth. Unfortunately, this L2 loss results in blurry images. This is so mainly because there are many possible candidates on a missing region given its context image. Since L2 loss aims to minimize pixel-wise errors, it produces the average out of distributions. On the other hand, the adversarial loss has been known to pick a particular mode and generate realistic images.

The context encoder uses the combination of the L2 loss and adversarial loss, to compute a reasonable prediction that also looks realistic. The adversarial loss is based on GAN (Generative Adversarial Networks), which consists of a generative model, G, and an adversarial discrimina-



(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

Figure 9.2: Human can fill or predict the missing region out of its context. We consider a context encoder performing such process in a deep learning framework. We test two different loss functions: L2 and Adversarial loss. L2 tends to produce blurry results, since L2 aims to generate the average pixel values. On the other hand, considering the adversarial loss as well as L2 shows realistic and sharper results. This image is excerpted from [PKD+16]

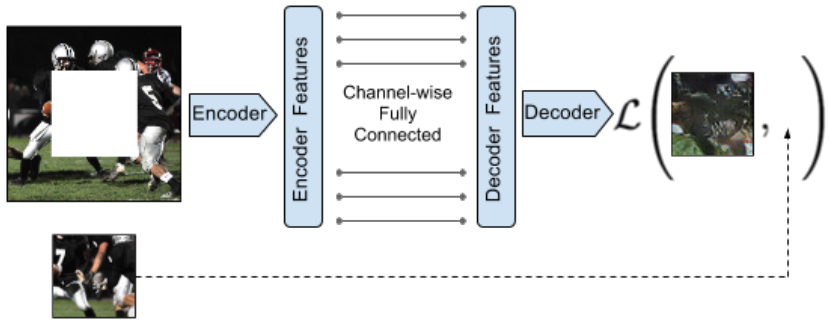


Figure 9.3: The context encoders shares the common structure of the autoencoder. This image is excerpted from [PKD+16]

tive model, D. Overall, GAN is based on a two-player game, where D takes a real image and a fake image generated from G and aims to discern whether the input image is real or fake. On the other hand, the generator aims to full D by creating more realistic images.

The objective function for the discriminator is based on the log likelihood indicating the input is real or fake:

$$\underset{G}{\operatorname{argmin}} \underset{D}{\operatorname{argmax}} E_x[\log(D(x))] + E_z[\log(1 - D(G(x)))], \tag{9.1}$$

where x and z are images from real and fake images, respectively. The context encoder adopts this framework and use the context encoder as the generator. The discriminator used in the context encoder shares similar architectures of the encoder itself.

The context encoder is tested on various applications other than the inpainting application. In the inpainting application (Fig. 9.2), we also get reasonable results even though we have large holes in the input images. Nonetheless, when there are many textures that we can utilize for filling holes, existing techniques based on texture synthesis tend to show better than the context encoder.

In additional to the inpainting process, we can use the context encoder to learn features in an supervised manner for image search. It can be also used for pre-training on various tasks such as classification and detection, and was demonstrate to work better than other unsurpervised pre-training methods (e.g., autoencoder).

9.3 Classification Pre-Training

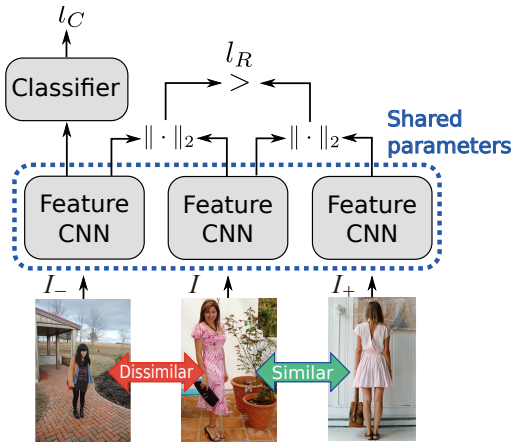


Figure 9.4: For utilizing weak labelled data, we use both classification and ranking losses. The ranking loss, l_R, is used for adjusting features, while the classification loss is used for updating features and classifier. This image is excerpted from [SSI16]

9.4 Utilizing Weak Labels

Some of labels, especially data acquired from internet, are very noisy and diverse. These data are treated as weak labels, compared to well organized datasets such as ImageNet. Training network with these noisy and weak labels may not result in high classification accuracy. Nonetheless, it is critical to utilize them well, since data with weak labels are more abundant compared to well organized and curated datasets. In this section, we discuss various approach of utilizing those data with weak labels.

Joint optimization. Simply training with a classification loss with weak labels may not lead to learn something useful. A remedy is to jointly learn features with the classification loss and ranking loss utilizing triplets (Fig. 9.4). Similarity among the triplets can be measured by their tags such as tag intersection divided by their union. The ranking loss does not affect the adopted classifier, but updates CNN features based on the similarity among the triplets. This simple joint optimization shows higher accuracy with a compact feature over only using a ranking or classification, and even over the Siamese network.

Bibliography

[Alp04] Ethem Alpaydin. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004.

[BaI08] O. Boiman and E. Shechtman andM. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[BETG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008.

[Boo89] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989.

[Bre01] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[CBHK02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, 2002.

[CCZH17] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *The Conference on Computer Vision and Pattern Recognition*, 2017.

[CHL12] Yan-Ying Chen, Winston H. Hsu, and Hong-Yuan Mark Liao. Discovering informative social subgraphs and predicting pairwise relationships from group photos. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 669–678, 2012.

[CLW+14] Jian Cheng, Cong Leng, Jiaxiang Wu, Hainan Cui, and Hanqing Lu. Fast and accurate image matching with cascade hashing for 3d reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[CZ07] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, pages 1 –8, 2007.

[CZLT14] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE CVPR*, 2014.

[DFG14] S.K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[DGE15] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[DT05] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893 vol. 1, 2005.

[ESTA14] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.

[eY17] Sung eui Yoon. *Rendering*. 2017. Available freely online.

[FGM10] P.F. Felzenszwalb, R.B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2241–2248, 2010.

[FMR08] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[GDDM14] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[GHKS13] Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[GKHE10] M. Grundmann, V. Kwatra, Mei Han, and I Essa. Efficient hierarchical graph-based video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2141–2148, 2010.

[GL11] Y. Gong and S. Lazebnik. Iterative quantization: a procrustean approach to learning binary codes. In *CVPR*, 2011.

[GRSPV12] A. Gordo, J.A. Rodriguez-Serrano, F. Perronnin, and E. Valveny. Leveraging category-level labels for instance-level image retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3045 –3052, june 2012.

[GSY+17] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 2017.

[GXY+17] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Trans. Graph.*, 2017.

[GZMT10] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[HAWL04] Chang Huang, Haizhou Al, Bo Wu, and Shihong Lao. Boosting nested cascade detector for multi-view face detection. 2004.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[HS81] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3), 1981.

[HS88] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50, 1988.

[JBL15] Justin Johnson, Lamberto Ballan, and Fei-Fei Li. Love thy neighbors: Image annotation by exploiting image metadata. *ICCV*, 2015.

[JDS11] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 2011.

[JDSP10] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304 –3311, 2010.

[KLK14] K. Karsch, C. Liu, and S. B. Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[KTS+14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LDPT17] Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graph.*, 2017.

[LF10] Ce Liu and William T. Freeman. *A High-Quality Video Denoising Algorithm Based on Reliable Motion Estimation*. 2010.

[Lin98] Tony Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[LK81] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. IJCAI'81, 1981.

[LLAH14] Keyu Lu, Jian Li, Xiangjing An, and Hangen He. A hierarchical approach for road detection. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 517–522, 2014.

[LLS13] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. In *Proceedings of Robotics: Science and Systems*, 2013.

[LM71] Edwin H Land and John J McCann. Lightness and retinex theory. *JOSA*, 61(1):1–11, 1971.

[LMSR08] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.

[LSP06] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[LTS+14] Xiao Liu, Dacheng Tao, Mingli Song, Ying Ruan, Chun Chen, and Jiajun Bu. Weakly supervised multiclass video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[LY15] Guanbin Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015.

[LYT11] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):978–994, 2011.

[MPK08] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. *Computer Vision – ECCV 2008*, chapter A New Baseline for Image Annotation. 2008.

[MS01] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *the International Conference on Computer Vision*, pages 525–531 vol.1, 2001.

[MS04] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[NF13] Mohammad Norouzi and D Fleet. Cartesian k-means. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[OW04] I. Omer and M. Werman. Color lines: image specific color representation. In *CVPR*, volume 2, 2004.

[PD07] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[PKD+16] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[PL11] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *the International Conference on Computer Vision*, 2011.

[PSGIN+16] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor. Shallow and deep convolutional networks for saliency prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[QGB+11] Danfeng Qin, S. Gammeter, L. Bossard, T. Quack, and L. van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR*, 2011.

[RAS17] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. *CoRR, CVPR 17*, 2017.

[RDSJ13] Jérôme Revaud, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Event retrieval in large video collections with circulant temporal encoding. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013.

[RHGS15] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, 2015.

[RSTK03] Jason Rennie, Lawrence Shih, Jaime Teevan, and David Karger. Tackling the poor assumptions of naive bayes text classifiers. In *In Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[Sah96] Mehran Sahami. Learning limited dependence bayesian classifiers. In *Knowledge Discovery and Data Mining*, pages 335–338, 1996.

[SGE12] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision*, 2012.

[SLB+12] Xiaohui Shen, Zhe Lin, J. Brandt, S. Avidan, and Ying Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[SLBW13] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Detecting and aligning faces by image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[SMGE11] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph.*, 30(6):154:1–154:10, 2011.

[SSI16] Edgar Simo-Serra and Hiroshi Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *CVPR*, 2016.

[STZ+16] Xiaoyong Shen, Xin Tao, Chao Zhou, Hongyun Gao, and Jiaya Jia. Foremost regional matching for internet scene images. *ACM Transactions on Graphics*, 2016.

[SWK16] Christoph Rhemann Shahram Izadi Shenlong Wang, Sean Fanello and Pushmeet Kohli. The global patch collider. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Sze10] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., 1st edition, 2010.

[Tai14] DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[UvdSGS13] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511–I–518 vol.1, 2001.

[Wei01] Yair Weiss. Deriving intrinsic images from image sequences. In *the International Conference on Computer Vision*, 2001.

[WGC12] Zhengxiang Wang, Shenghua Gao, and Liang-Tien Chia. Learning class-to-image distance via large margin and l1-norm regularization. In *ECCV*, Lecture Notes in Computer Science, pages 230–244. 2012.

[WGH14] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3302–3309, 2014.

[WKC10] J. Wang, S. Kumar, and S.-F. Chang. Semi-supervised hashing for scalable image retrieval. In *CVPR*, 2010.

[WRHS13] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. DeepFlow: Large displacement optical flow with deep matching. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[WYZL13] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *the International Conference on Computer Vision*, 2013.

[yCH09] Ming yu Chen and Alex Hauptmann. Mosift: Recognizing human actions in surveillance video. Technical report, CMU, 2009.

[YSRL11] J. Yagnik, D. Strelow, D.A. Ross, and Ruei-Sung Lin. The power of comparative reasoning. In *the International Conference on Computer Vision*, pages 2431–2438, 2011.

[YYGH09] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[Zha04] Harry Zhang. The optimality of naive bayes. *A A*, 1(2):3, 2004.

[ZLYM13] Guang-Tong Zhou, Tian Lan, Weilong Yang, and Greg Mori. Learning class-to-image distance with object matchings. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.