

Simulating vision through time: Hierarchical, sparse models of visual cortex for motion imagery

Amy E. Galbraith, Steven P. Brumby, and Rick Chartrand
Los Alamos National Laboratory
Mailstop B244, Los Alamos, NM 87545
Email: amyg@lanl.gov, brumby@lanl.gov, rickc@lanl.gov

Abstract—Efficient pattern recognition in motion imagery has become a growing challenge as the number of video sources proliferates worldwide. Historically, automated analysis of motion imagery, such as object detection, classification and tracking, has been accomplished using hand-designed feature detectors. Though useful, these feature detectors are not easily extended to new data sets or new target categories since they are often task specific, and typically require substantial effort to design. Rather than hand-designing filters, recent advances in the field of image processing have resulted in a theoretical framework of sparse, hierarchical, learned representations that can describe video data of natural scenes at many spatial and temporal scales and many levels of object complexity. These sparse, hierarchical models learn the information content of imagery and video from the data itself and lead to state-of-the-art performance and more efficient processing. Processing efficiency is important as it allows scaling up of research to work with dataset sizes and numbers of categories approaching real-world conditions. We now describe recent work at Los Alamos National Laboratory developing hierarchical sparse learning computer vision models that can process high definition color video in real time. We present preliminary results extending our prior work on object classification in still imagery [1] to discovery of useful features at different time scales in motion imagery for detection, classification and tracking of objects.

I. INTRODUCTION

Vision is one of the hardest and most intriguing problems in artificial intelligence and computer science, and automated human-like annotation of video is a key enabling technology for large-scale data retrieval and search applications. Advances in this field face three fundamental challenges: (1) how to create mathematical models of natural video sequences that can reconstruct, interpret and predict the visual scene; (2) how to learn image features across many spatial and temporal scales for many object categories; and, (3) how to exploit the sheer size and richness of large-scale video datasets now becoming available.

We report on recent and ongoing work to extend the theoretical framework of sparse representations to learn hierarchical representations that describe video data of natural scenes at many spatial and temporal scales and many levels of object complexity. A typical frame of video that we would like to parse for multi-object classification is shown in Fig. 1. Such a scene contains hundreds of discrete objects that fall in a natural hierarchy (e.g., a bus on a road in a city). We wish to understand the types of algorithms and computing platforms necessary to detect and classify multiple objects in video,



Fig. 1. Any given frame of video of a natural scene can contain hundreds of objects drawn from tens of thousands of object categories. The human eye and visual cortex in brain processes petapixels of video data per year. This prompts the research question: what types of algorithms can scale to detect, classify and track objects at this level of data volume and scene complexity?

including higher-level learning of actions and interactions of moving objects. Detection of actions and behavior in video is essential to automatic scene understanding and is a growing area of interest. The amount of new video data (for instance, more than 70 hours of video are uploaded to YouTube every minute [2]) far exceeds the resources for humans to do manual markup and description of activities and behaviors. Interesting motion happens at many different time scales: less than one second for part-based motion (moving limbs) or texture (water splashing in a fountain), to multiple seconds (person interactions), to tens of seconds (bicycle traversing a path), to minutes or hours (time-lapse photography).

The standard approach for learning the visual appearance of multiple objects in video is to break the video sequence into a set of patches of fixed spatial size and number of frames (i.e., localized in both space and time) and then extract a set of features characterizing those patches. The patches may be extracted at multiple scales, but analysis of global structure in video has proven difficult and so many researchers have adopted a *bag-of-words* approach, representing an image by the histogram of local feature responses. State-of-the-art results for multi-category object recognition have used matching of spatial pyramids [3], [4] of histograms of oriented Gabor features [5], patch averages of orientation and texture

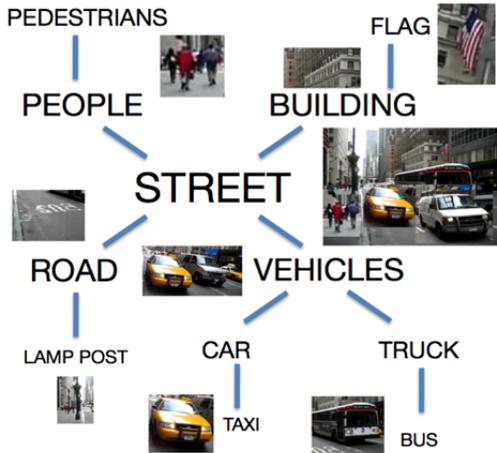


Fig. 2. Manual hierarchy of object categories in Fig. 1. Automatic annotation of large, rich imagery and video datasets enables search, retrieval and analysis applications.

features [6], or neuroscience-inspired vectors of local texture, color and motion features [7], [8].

More recently, investigations of the statistics of natural scenes have suggested that natural signals are sparse, i.e., they can be represented by only a few terms from an over-complete "dictionary" (generalized basis set) of signal "atoms". Donoho and many others have exploited this structure to produce state-of-the-art algorithms for signal compression, denoising, signal reconstruction, super-resolution, and inpainting [9], [10], [11]. The dictionary can be pre-specified (e.g., some wavelet packet dictionary), or can be learned from the data itself [12]. Very recently, research has turned to problems of tuning sparse representations for object classification [13]. For a single-layer, single-scale representation, the task is to find a dictionary Φ and coefficient vector S such that S is as sparse as possible and ΦS approximates the signal X sufficiently well. However, even for a fixed dictionary, the optimization problem of finding the sparsest S giving a tolerable amount of approximation error is computationally intractable to solve directly, being an NP-hard problem. Under many circumstances, however, a convex surrogate problem which can be solved efficiently recovers the sparsest solution [9], [14]. This discovery has led to many new algorithms and theoretical advances for sparse representation, with recent progress driven particularly by mathematically-equivalent problems in compressive sensing [15], [16]. We have previously shown [17], [18] that solving a non-convex problem instead dramatically outperforms convex methods in theory and in practice, and can be implemented in an efficient way that does not encounter problems with local minima [19]. This can loosen the requirements for successfully computing a sparse representation, such as in information-rich contexts where the sparsest possible representation has twice as many nonzero components as that for which a convex method could successfully solve.

In this paper, we present a framework to exploit the rich, complementary features in video. We propose combining local

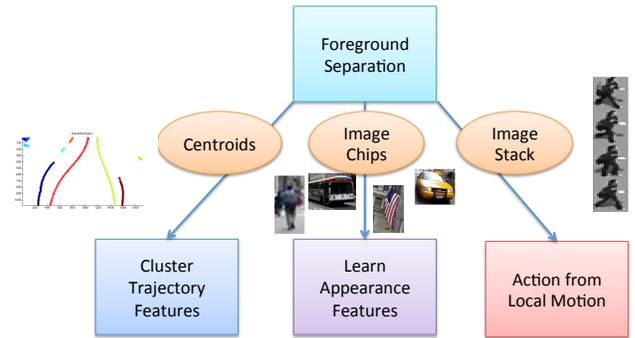


Fig. 3. Multiple complimentary features used for detection, classification, and tracking of moving objects in video.

motion, local appearance, and global motion or trajectory features, to detect, classify and track objects in video. Our approach is inspired by recent advances in behavior analysis of animal subjects [20], [21] and group activity analysis [22]. However, our foreground separation algorithm [23] is an improvement upon similar methods and greatly simplifies the learning of appearance and tracking of the objects of interest. We add a nonconvex regularization and apply a splitting approach to decompose the optimization problem into simple, parallelizable components. The nonconvex regularization supports our behavior and appearance learning tasks, as it preserves shapes and contrast better than prior methods [24]. Figure 3 illustrates the processing paths used to demonstrate multiple-object classification. We discuss our non-convex robust PCA foreground separation in Sect. I-B, which can provide a clean pixel-based mask for analysis of appearance (Sect. I-C) and trajectory (Sect. I-D). This work extends previous research on appearance-based classification alone [1]. We also briefly discuss how the same framework can support action recognition algorithms in Sect. I-E.

A. Data

This work uses a dataset provided by the DARPA/DSO NeoVision2 Program. Here we use the Tower dataset collected by Sebastian Thrun's research team at Stanford. This dataset consists of 30 second clips of high definition (1080p) color video recorded from a camera with fixed orientation located in a multi-story building overlooking a traffic circle through which passes a stream of pedestrians, cyclists and motor vehicles, as shown in Fig. 4. Lighting conditions varied throughout the day of data collection, producing varying shadows, and a light wind produced moderate levels of canopy movement in surrounding vegetation. The complete Tower dataset consists of several hours of video of this kind, for which ground truth is available in the form of frame-by-frame manual mark-up for ten categories of foreground objects.

B. Non-Convex Robust PCA Foreground Separation

Principal component analysis (PCA) is a widely used data analysis tool that finds the smallest set of orthogonal basis vectors on which the columns of a data matrix D can be



Fig. 4. Example video frame: High definition 1080p color video frame from DARPA NeoVision2 dataset Tower 014. Foreground objects include pedestrians, cyclists and cars.

represented to a specified accuracy. While PCA is effective in the presence of Gaussian white noise, it can give very poor results when even a small subset of the entries in matrix D are corrupted by outliers, such as in the case of impulse noise. A variant of PCA that is robust to a sparse set of corrupted data values can be constructed as the solution to the optimization problem,

$$\min_{L,S} \text{rank}(L) + \lambda \|S\|_0 \text{ such that } L + S = D, \quad (1)$$

where $\|\cdot\|_0$ counts the number of nonzero entries, and $\lambda > 0$ is a tuning parameter. We can regard L as a low-rank approximation to D , and S as a sparse set of possibly large deviations from that approximation. While this optimization problem is intractable, Candès *et al.* [25] have proposed the tractable, convex relaxation

$$\min_{L,S} \|\sigma(L)\|_1 + \lambda \|S\|_1 \text{ such that } L + S = D, \quad (2)$$

where $\|S\|_1$ is the ℓ^1 norm of S (the sum of the absolute values of all entries in S), and $\sigma(L)$ is the vector of singular values of L , with $\|\sigma(L)\|_1$ also known as the *nuclear norm* of L .

While this method has been considered for a variety of applications, of specific interest here is the application to automated background removal in video. This is done by applying (2) to the matrix D having each column be a (vectorized) frame of a video clip. This approach has been shown to provide very good performance [26]. (It is worth noting, however, that there is a significant body of prior approaches to constructing robust variants of PCA, including applications to video background modeling [27].) The reason for this is that motion of objects within an image is a highly nonlinear process: every new pixel occupied by a moving object represents an independent dimension that the object's trajectory occupies in pixel space. Since L represents an approximation of D by a linear subspace whose dimension is the rank of L , the video represented by L can contain very little motion if the rank of L is to be low. Also note that extending the model of [25] to include a total variation (TV) regularization term has been shown to improve performance of this approach in video background modeling [28], [23]. A

number of fast algorithms [27], [29], [30], [31], [32], [33], [34], [35] have been proposed to avoid the computational expense of directly solving (Eq. (2)), but these appear to trade separation quality for speed.

The robust PCA model of (2) can be improved by replacing the ℓ^1 norms by *nonconvex* penalty functions that promote sparsity more strongly. This is in analogy with the field of compressive sensing [15], [16], where using the ℓ^p quasi-norm instead (defined by $\|x\|_p^p = \sum_i |x_i|^p$) with $p < 1$ gives better results, such as robustness to noise and signal nonsparsity [36], [18], [37]. In the case of robust PCA for background subtraction, a nonconvex approach can tolerate more pixels in the sparse component S , typically produces a component L with much lower rank, and is better able to separate objects that are both moving and stationary at different portions of the same clip [23].

Our approach is to solve the following modification of (2):

$$\min_{L,S} \|\sigma(L)\|_p + \lambda \|S\|_p \text{ such that } L + S = D. \quad (3)$$

Our penalty function $\|\cdot\|_p$ is a modification of the ℓ^p norm, designed to be very efficient to minimize and to work well with an alternating direction, method-of-multipliers algorithm (also known as split Bregman). The key idea is that the solution to the optimization problem

$$\min_X \|X\|_p + \frac{1}{2\mu} \|X - Y\|_2^2 \quad (4)$$

is given by a *shrinkage* operation, a generalization of soft thresholding:

$$X_{ij} = \max\{0, |Y_{ij}| - \mu|Y_{ij}|^{p-1}\} \frac{Y_{ij}}{|Y_{ij}|}. \quad (5)$$

Note that when $p = 1$, (5) is precisely soft thresholding, reflecting the fact that $\|\cdot\|_p$ reduces to the ℓ^1 norm when $p = 1$. Further details regarding $\|\cdot\|_p$ can be found in [23].

We applied our method to several video clips of the Tower data with multiple moving objects including pedestrians, cyclists, vehicles, a fountain, and trees moving in the wind. The sparse component S contains the moving objects within the video, mainly pedestrians and cyclists in the example shown, as seen in Fig. 5. As shown in Fig. 6, the low rank component L is primarily the stationary background, but can show objects that have stopped moving (such as the pedestrian at the bottom edge of the fountain), as well as slow variations in the background such as the overall lighting conditions. Object detection is carried out on the sparse component S by thresholding on pixel values in S and then running a simple connected components algorithm on the selected pixels. Connected components are accepted as a detection if the component contains at least 150 pixels.

C. Appearance Based Classification Model

Once candidate foreground regions have been detected and segmented, we seek to automatically filter out false alarms, e.g., motion in swaying foliage or a splashing fountain, using appearance based classification of individual segmented

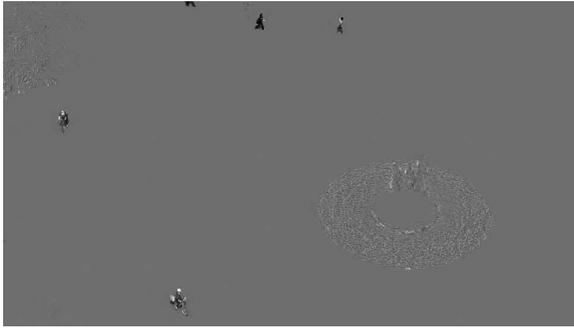


Fig. 5. Sparse component of video for a single frame. The moving foreground objects are well-separated from the uniform background.



Fig. 6. Low rank component of video for a single frame.

regions (ignoring temporal cues for the moment). Identifying image segments under real-world conditions of many, heterogeneous object categories (e.g., pedestrians, cyclists, vehicles) and uncontrolled illumination conditions is a very active area of research. Further, human visual system processes the equivalent of over a petapixel of video imagery per year, and the computer vision research community has started to explore the scaling of computer vision models to thousands of object categories in terapixel-sized datasets [38].

Our approach to this problem is to explore a class of neuroscience-inspired computer vision algorithms for object classification. Observations of sparsely activated neurons in visual cortex have led to computer vision algorithms based on sparse image-patch representations using adaptive, over-complete image feature dictionaries learned from data [39]. These models are generative extensions of the HMAX model [8], allowing reconstruction of the input image. They can also drive many-category classification of image patches for object detection within a large video frame. Building models that can simulate full-scale visual systems (primary visual cortex V1 alone contains hundreds of millions of neurons) and experiment with large datasets requires use of high performance computing platforms, and we are developing a fast, parallel implementation of these algorithms, called PANN [1].

PANN is a multi-layer convolutional network model that learns a sparsifying over-complete color/texture feature dictionary for the dataset, Figs. 7,8. Our first layer (modeled loosely on primary visual cortex (V1) S-cell layer) uses a learned

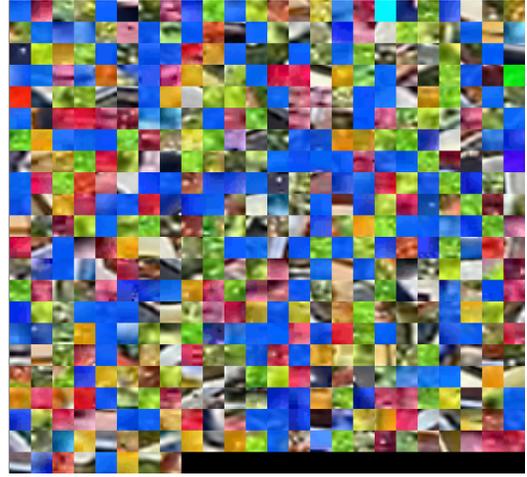


Fig. 7. Background Dictionary: 512 element dictionary trained using 11x11 pixel color patches from the video background.



Fig. 8. Foreground Dictionary: 512 element Dictionary trained using 11x11 pixel color patches from the extracted foreground segments (people and cyclists).

dictionary Φ to build a local sparse representation S for data X using a matching pursuit algorithm,

$$\min_{S, \Phi} \|X - \Phi S\|_2^2 + \lambda \|S\|_0 . \quad (6)$$

Our representations are very sparse, with typically less than 5% of local feature detectors active in any given column. However, they still allow for good reconstruction of the input image, in distinction to standard HMAX approaches that are feed-forward systems only. Later layers in the model cluster the patches in this high dimensional but sparse feature space producing a local image patch descriptor. These local descriptors can be pooled using, e.g., spatial pyramids [3] and passed to a support vector machine (SVM) [40]. We use the standard LIBSVM package by Chang, et al., [41]. Note that PANN analyzes each frame individually.

Previously, we applied PANN to DARPA NeoVision2 Tower and Helicopter video datasets to localize and classify objects



Fig. 9. Ground Truth and PANN’s appearance-based classification of detected segments for a training frame: ground truth boxes (orange), sparse connected-components (blue), and predicted foreground hits (red).



Fig. 10. Ground Truth and PANN’s appearance-based classification of detected segments for a test frame: ground truth boxes (orange), sparse connected-components (blue), and predicted foreground hits (red).

by searching the full frame [42], [43]. However, when applied to multiple frames, the detections show erroneous variation from frame to frame, indicating the importance of ensuring temporal consistency. Detection of temporal features to improve detections using PANN is discussed in Sect. I-D.

For the current work, we applied PANN to the sparse piece of our low rank plus sparse decomposition. We considered bounding boxes of connected regions of the sparse component that meet a threshold on size. For these patches, we train PANN to learn a dictionary, and use LIBSVM to train a linear kernel support vector machine classifier to filter detections of people and cyclists from “false alarm” detections (e.g., trees, fountain, shadows). This turns out to be a easy task now that the low rank plus sparse decomposition has localized the salient regions of the frame. We used ground truth provided by the DARPA NeoVision2 program to label a set of 1295 sparse component patches for training, and an independent test set of an additional 881 patches for our test set. We achieve a classification accuracy of 98% on our test set (866/881), with an example result for a specific frame shown in Figs. 9,10. Note that this accuracy is with respect to moving objects. Target objects that remain stationary through the majority of the video sequence (e.g., parked cars) are not detected by the low rank plus sparse decomposition of the video sequence, and hence are not available as input to our segment classifier.

D. Trajectory Features

After computing regions of interest using nonconvex foreground separation, we can extract trajectories in order to classify behavior. Because the sparse frames have a background set to a constant, we can use a simple threshold to generate a binary mask of moving objects. The choice of threshold is in practice not very sensitive for the relatively clean Tower videos, and is simply chosen to be a reasonable distance away from the background (with background set as the mean of the frame for simplicity). Since the output of the decomposition was in 16-bit signed integer values, we used a two-sided threshold, looking for those values s such that $|s - \frac{2^N}{2}| > \mu$ exceeds some value μ , where N is the bit depth of the image frames for signed integer values; in the video clips here, we

used $\mu = 5000$, or approximately 7.6% of the dynamic range of the sparse foreground image.

After binarization, a 2D connected components algorithm was used to compute a centroid for each object. Objects smaller than 150 pixels were discarded. A particle filter was applied to the centroids, resulting in linked tracks. Particle filters for computing trajectories have been very successful in recent years; we used the particle filter implementation of Blair and Dufrense [44], based upon the popular IDL implementation of Crocker and Grier [45], [46]. Although we used a very basic implementation here, efficient implementations of particle filters (including on GPU) for realtime tracking have been developed [47], [48], [49] that could be used to accelerate tracking for our goal of realtime trajectory analysis.

Illustrative results to compute centroids and link them into persistent tracks are shown in Fig. 11, with quantitative results compared to hand annotated ground truth tracks shown in Table I, using another video from the DARPA Tower dataset. This particular video was a zoomed region of the overall scene that contained tree motion and occlusions as well as pedestrians and bicycles. We discarded a small number of very short tracks and focused on the persistent tracks, since we were mainly concerned with analysis of significant actors in the scene. With a frame rate of 30 fps, tracks less than one second in length are not likely to be human-caused. However, we wished to distinguish significant, persistent motion from the trees from human activity. For a small number of positions, the ground truth annotations were incomplete, meaning that a visual inspection of the video indicated the presence of an object that was missing in the provided bounding box list. Therefore, we augmented the ground truth via visual inspection to ensure that our detected track positions were not greater in number than those of the corresponding ground truth positions.

Using the detected tracks, we demonstrate the potential of our foreground separation for performing group activity analysis. We computed simple velocity and direction features using the track positions and times, and did a simple clustering by the mean velocity and direction, which easily separated the random direction of the canopy movement from

TABLE I
DETECTED TRACK POSITIONS VERSUS GROUND TRUTH.

Detections	Truth	Percent
265	270	98.1
44	30	100
51	61	83.6
142	121	100
214	211	100
247	271	91.1
61	181	33.7



Fig. 11. The ten longest tracks in zoomed tower video. Note occlusions of sidewalk by vegetation leading to loss of track persistence at the particle filtering step.

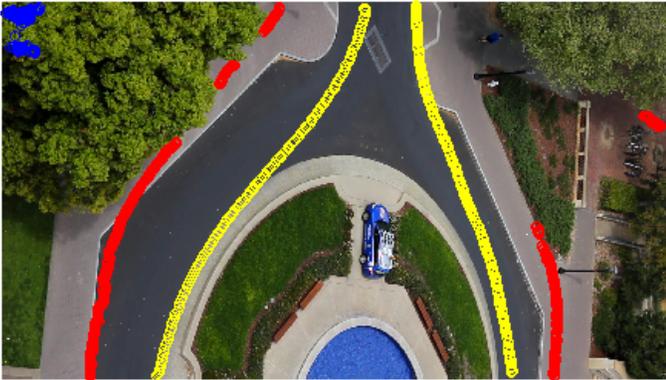


Fig. 12. Clustering of tracks by velocity and heading to demonstrate group behavior.

the consistent direction of the human activities, as shown in Fig. 12. Smoothed velocities of the pedestrians and bicycles plotted along with ground truth are shown in Fig. 13. The velocities are more inconsistent in areas of canopy occlusion and are an area of future work. More sophisticated trajectory clustering is becoming a large field of research due to its diverse potential applications. A recent overview of trajectory clustering for activity analysis in video is given in [50]. A main area of investigation is the selection of an appropriate metric to measure the similarity of trajectory segments, as well as the ability for these techniques to be applied to difficult real-world applications [51], [52].

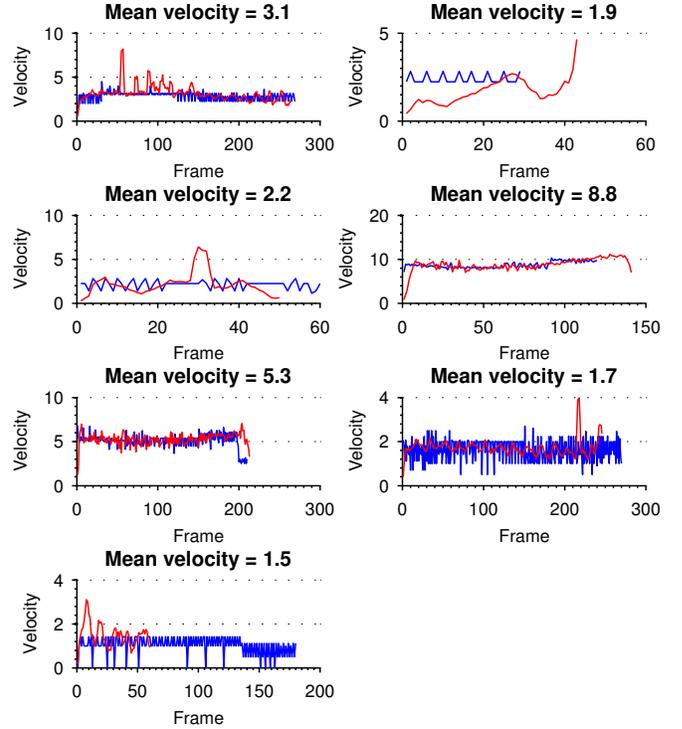


Fig. 13. Smoothed track velocities for pedestrians and bicycles (red) and ground truth velocities (blue). A four-point moving average was used.

E. Local Motion Features

In addition to learning appearance models and computing trajectories, there is a growing interest in the area of local action recognition. Action recognition has reached a point of maturity for simple test databases such as KTH [53], which typically contain a limited set of actions in a constrained setting. Extracting motion out of video in unconstrained environments is the first step in extending action recognition to real world problems. Our low rank plus sparse decomposition of video frames can be used as a preprocessing step for action recognition due to the very clean pixel-by-pixel boundaries extracted for moving objects. We took the centroids from the connected components step and simply set the bounding box size to that of the first bounding box detected in a trajectory, plus a few pixels for padding. Examples of the clean, sparse extracted regions of interest are shown in Fig. 14. Without the cluttered background, the local motion of a person walking and a bicyclist pedaling is readily apparent and may be used in a number of action recognition methods to cluster activity based upon cyclic motion of the actor throughout its trajectory. In addition, the change of action over time may be discovered by doing action recognition over sliding or disjoint time segments. As an example, a bicyclist may dismount a bike and then be more properly classified as a pedestrian. Therefore, along with learning appearance and computing trajectories, local action may provide complementary features for our multi-object classification approach and is made feasible using clean

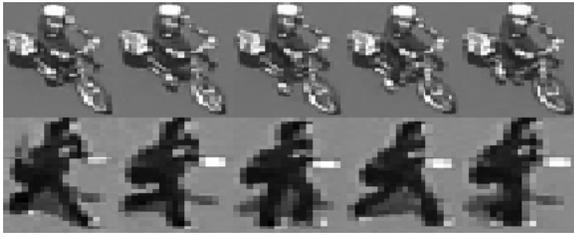


Fig. 14. Bicycle and pedestrian extracted from centroids in four consecutive sparse foreground frames.

foreground separation techniques, such as our low rank plus sparse decomposition.

II. DISCUSSION AND FUTURE WORK

Sparse signal processing techniques have been successfully applied to many areas of image and video analysis, and are just starting to be applied to real-world datasets. Applying these algorithms to high definition video collected in real-world settings, with unconstrained illumination and scene content, presents an interesting opportunity to study the performance of these algorithms in the limit of large volumes of data and large number of object categories. The current work has shown how sparse signal processing techniques developed separately for foreground extraction from video and for object classification in still imagery, can be combined and used to develop models of object trajectories for tracking and classification. New nonconvex methods can improve foreground extraction by decreasing the rank of the low rank component in low rank plus sparse decomposition of video. Sparse generative models can provide a useful approach to classification of candidate objects extracted from the sparse component of a low rank plus sparse video decomposition. Trajectory based classification can be carried out independently from and complementary to appearance based classification, and provides a way of imposing temporal consistency on frame-by-frame appearance-based classification. Further, trajectory analysis is one promising way to get to timescales of one to a few seconds, equivalent to hundreds of frames of video, at which many human activities may be defined.

In future work, we will consider how appearance-based classification can help identify fragmented trajectories caused by, e.g., obscuration from tree canopies or buildings, and will consider the scaling behavior as we increase the number of foreground and background classes.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the Department of Energy through the LANL/LDRD Program. The authors thank the DARPA/DSO NeoVision2 Program and Prof. Thrun's Lab at Stanford for use of the NeoVision2 Tower dataset. The authors would like to thank Brendt Wohlberg for useful discussions.

REFERENCES

- [1] S. Brumby, G. Kenyon, W. Landecker, C. Rasmussen, S. Swaminarayan, and L. Bettencourt, "Large-scale functional models of visual cortex for remote sensing," in *Applied Imagery Pattern Recognition Workshop (AIPRW), 2009 IEEE*, Oct. 2009, pp. 1–6.
- [2] [Online]. Available: http://www.youtube.com/t/press_statistics
- [3] C. S. S. Lazebnik and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition (CVPR)*, 2006.
- [4] J. Lasserre, C. Bishop, and T. Minka, "Principled hybrids of generative and discriminative models," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, June 2006, pp. 87–94.
- [5] D. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2, 1999, pp. 1150–1157.
- [6] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.
- [7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
- [8] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Science*, vol. 104, pp. 6424–6429, April 2007.
- [9] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2197–2202, 2003.
- [10] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010.
- [11] B. Wohlberg, "Inpainting by joint optimization of linear combinations of exemplars," *Signal Processing Letters, IEEE*, vol. 18, no. 1, pp. 75–78, Jan. 2011.
- [12] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, June 1996.
- [13] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 4, pp. 791–804, Apr. 2012.
- [14] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
- [15] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, pp. 489–509, 2006.
- [16] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1289–1306, 2006.
- [17] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.
- [18] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 035020, pp. 1–14, 2008.
- [19] R. Chartrand, "Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data," in *IEEE International Symposium on Biomedical Imaging*, 2009.
- [20] X. Burgos-Artizzu, P. Dollar, L. Dayu, and D. J. Anderson, "Social behavior recognition in continuous video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [21] H. Jhuang, E. Garrote, X. Yu, V. Khilnani, T. Poggio, A. Steele, and T. Serre, "Automated home-cage behavioural phenotyping of mice," *Nature Communications*, Sep. 2010.
- [22] Z. Tang, A. Castrodad, M. Tepper, and G. Sapiro, "Are you imitating me? unsupervised sparse modeling for group activity analysis from a single video," *arXiv preprint arXiv:1208.5451*, 2012.
- [23] R. Chartrand, "Nonconvex splitting for regularized low-rank + sparse decomposition," *Signal Processing, IEEE Transactions on*, vol. 60, no. 11, pp. 5810–5819, Nov. 2012.
- [24] —, "Nonconvex regularization for shape preservation," in *IEEE International Conference on Image Processing*, 2007.

- [25] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, pp. 11:1–11:37, June 2011.
- [26] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Adv. in Neural Inf. Proc. Sys. (NIPS) 22*, 2009, pp. 2080–2088.
- [27] Y. Li, "On incremental and robust subspace learning," *Pattern Recognition*, vol. 37, no. 7, pp. 1509–1518, 2004.
- [28] B. Wohlberg, R. Chartrand, and J. Theiler, "Local principal component pursuit for nonlinear datasets," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012.
- [29] G. Mateos and G. Giannakis, "Sparsity control for robust principal component analysis," in *Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov. 2010, pp. 1925–1929.
- [30] C. Qiu and N. Vaswani, "Real-time robust principal components' pursuit," *CoRR*, vol. abs/1010.0608, 2010.
- [31] —, "Reprocs: A missing link between recursive robust PCA and recursive sparse recovery in large but correlated noise," *CoRR*, vol. abs/1106.3286, 2011.
- [32] G. Pope, M. Baumann, C. Studer, and G. Durisi, "Real-time principal component pursuit," in *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Nov. 2011, pp. 1433–1437.
- [33] R. Liu, Z. Lin, S. Wei, and Z. Su, "Solving principal component pursuit in linear time via l_1 filtering," *CoRR*, vol. abs/1108.5359, 2011.
- [34] T. Zhou and D. Tao, "Godec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, June 2011, pp. 33–40.
- [35] J. He, L. Balzano, and A. Szelam, "Incremental gradient on the grassmannian for online foreground and background separation in subsampled video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1568–1575.
- [36] R. Chartrand, "Exact reconstructions of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, pp. 707–710, 2007.
- [37] R. Saab, R. Chartrand, and Özgür Yilmaz, "Stable sparse approximations via nonconvex optimization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.
- [38] J. Deng, A. Berg, K. Li, and L. Fei-Fei, "What does classifying more than 10,000 image categories tell us?" in *Proceedings of the 12th European Conference of Computer Vision (ECCV)*, 2010.
- [39] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
- [40] V. V. Corinna Cortes, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
- [41] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [42] S. Brumby, M. Ham, and G. Kenyon, "Semi-supervised learning of high-level representations of natural video sequences," in *Computational and Systems Neuroscience (COSYNE)*, 23-26 Feb 2012.
- [43] D. Paiton, S. Brumby, G. Kenyon, G. Kunde, K. Peterson, M. Ham, P. Schultz, and J. George, "Combining multiple visual processing streams for locating and classifying objects in video," in *Image Analysis and Interpretation (SSIAI), 2012 IEEE Southwest Symposium on*, april 2012, pp. 49–52.
- [44] [Online]. Available: <http://physics.georgetown.edu/matlab/>
- [45] J. Crocker, D. Grier *et al.*, "Methods of digital video microscopy for colloidal studies," *Journal of colloid and interface science*, vol. 179, no. 1, pp. 298–310, 1996.
- [46] [Online]. Available: <http://www.physics.emory.edu/~weeks/idl/>
- [47] M. Goodrum, M. Trotter, A. Aksel, S. Acton, and K. Skadron, "Parallelization of particle filter algorithms," in *Computer Architecture*, ser. Lecture Notes in Computer Science, A. Varbanescu, A. Molnos, and R. van Nieuwpoort, Eds. Springer Berlin / Heidelberg, 2012, vol. 6161, pp. 139–149.
- [48] J. Brown and D. Capson, "A framework for 3d model-based visual tracking using a gpu-accelerated particle filter," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 1, pp. 68–80, jan. 2012.
- [49] O. Lozano and K. Otsuka, "Real-time visual tracker by stream processing," *Journal of Signal Processing Systems*, vol. 57, pp. 285–295, 2009.
- [50] B. Morris and M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 11, pp. 2287–2301, Nov. 2011.
- [51] B. Han, L. Liu, and E. Omiecinski, "Neat: Road network aware trajectory clustering," in *Distributed Computing Systems (ICDCS), 2012 IEEE 32nd International Conference on*, June 2012, pp. 142–151.
- [52] M. Gariel, A. Srivastava, and E. Feron, "Trajectory clustering and an application to airspace monitoring," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 4, pp. 1511–1524, Dec. 2011.
- [53] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, Aug. 2004, pp. 32–36 Vol.3.