Learning Sparse Prototypes for Crowd Perception via Ensemble Coding Mechanisms

Yanhao Zhang[1,2], Shengping Zhang[1], Qingming Huang[2], and Thomas Serre[1]

[1] Department of Cognitive Linguistic & Psychological Sciences
Institute for Brain Sciences
Brown University, Providence, 02912, USA
[2] School of Computer Science
Harbin Institute of Technology, Harbin, 150001, China
{yhzhang,qmhuang}@hit.edu.cn, {shengping_zhang,thomas_serre}@brown.edu

Abstract. Recent work in cognitive psychology suggests that crowd perception may be based on pre-attentive ensemble coding mechanisms consistent with feedforward hierarchical models of visual processing. Here, we extend a biological model of motion processing with a new dictionary learning method tailored for crowd perception. Our approach uses a sparse coding model to learn crowd prototypes. Ensemble coding mechanisms are implemented via structural and local coherence constraints. We evaluate the proposed method on multiple crowd perception problems from collective or abnormal crowd detection to tracking individuals in crowded scenes. Experimental results on crowd datasets demonstrate competitive results on par or better than state-of-the-art approaches.

Keywords: Sparse coding, Crowd perception, Biological vision.

1 Introduction

The perception of crowd behavior has become a popular area of study straddling multiple disciplines from cognitive psychology to computer vision. Over the years, several computer vision approaches to crowd perception have been proposed, drawing inspiration from disparate fields from sociology [1] to physics [2].

The so-called "social models" aim at characterizing the interaction between individuals in a crowd. This can be done explicitly using either systems of non-linear coupled equations as in the "social force" model [3,4] or implicitly via dynamic space-time correlations [5]. More recent work has extended some of these ideas using visual saliency [6], conditional random fields [7] or other energy-based approaches [8,9]. A measure of intended motion using space-time statistics [10] has been proposed as a model of people's "efficiency". The "collectiveness" of crowd scenes has been estimated using tools borrowed from machine learning including manifold-based similarity measures [11].

Representative physics-based approaches include methods based on chaotic invariants to represent people's trajectories [12] and methods based on stability analysis to identify different patterns of behaviors [13].

H.S. Park et al. (Eds.): HBU 2014, LNCS 8749, pp. 86–100, 2014.

Approaches borrowed from computational linguistic have also been applied to crowd perception including latent models [14,15] as well as bag-of-word and other related approaches for learning spatio-temporal occurrences of crowd motion patterns [16,17,18]. A notable approach based on the structural flow of scenes has been proposed in [19] and an approach for learning typical prototypes from correlations in atomic activities in [20].

Recently, several dictionary learning approaches have been proposed for learning crowd prototypes using sparse coding techniques [21,22,23] or closely related linear programming or matrix factorization techniques [20,24]. For instance, Lu et al. propose an efficient sparse coding approach for learning combinations of basis functions to detect abnormal events from pyramid video structures [22]. One of the main limitations with these methods is that they typically focus on modeling local motion patterns when patterns of crowd behavior tend to be more global. This leads to crowd representations that tend to be relatively unstable over time and fail to capture typical crowd peculiarities. This poses a challenge for applications ranging from the tracking of individuals in crowd to the recognition of crowd behaviors over long time periods.

Proposed approach and related work: Here, we investigate novel coding mechanisms to extend existing dictionary learning approaches with the aim to better capture the structural and collective characteristics of crowds. Our approach is motivated by recent developments in cognitive psychology, where it has been suggested that crowd perception may rely on pre-attentive ensemble coding mechanisms [25]. Human observers estimate the intended direction of briefly presented crowds of point-light walkers better than individual walkers. Such results have been taken as suggestive evidence that observers rapidly pool information from multiple walkers to estimate the movement of a crowd, very much in the spirit of feedforward hierarchical models (see [26,27] for reviews).

Feedforward hierarchical models of visual processing (including computational models of the visual cortex [28] and closely related convolutional networks [29]) have been shown to exhibit competitive performance for the recognition of individual human or animal activities. We propose a significant extension of a feedforward hierarchical model of the visual cortex [28] from the recognition of individual behaviors to group behaviors. Direction-selective motion units based on optical flow calculations are used as an input stage. Crowd prototypes are learned in intermediate stages of the motion processing hierarchy based on a sparse coding model. The proposed optimization learns crowd prototypes through ensemble coding mechanisms by jointly enforcing local structure and coherence in the input motion patterns.

Most closely related to our learning framework are learning approaches described above based on spatio-temporal representations of crowd motion patterns [5,15,30,31], structural flow [19] and correlations between atomic activities [20,11]. Compared to the original hierarchical model of motion processing [28], we show that the resulting learned prototypes are more selective and more easily interpretable. The overall hierarchical architecture leads to a compact visual

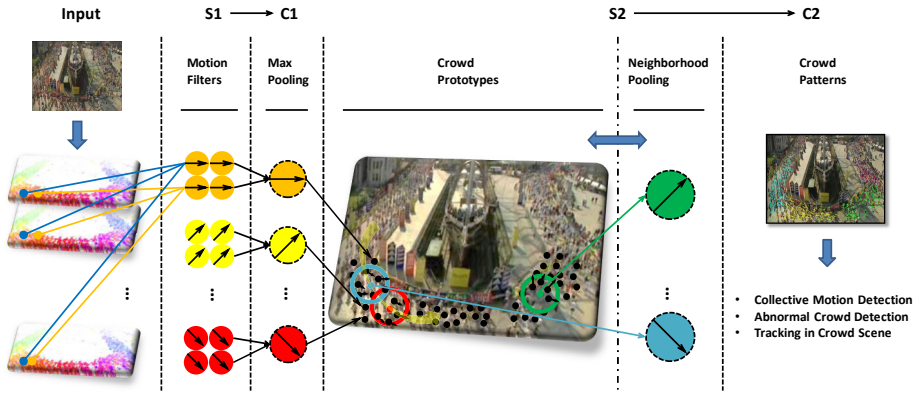


Fig. 1. Sketch of the proposed hierarchal model for crowd perception

representation capable of capturing the complex structure of motion patterns associated with crowd patterns.

In summary, this paper makes the following contributions: (1) We describe a novel mid-level representation together with an algorithm for learning crowd prototypes within a feedforward hierarchical model of motion processing; (2) motivated by biological considerations, we incorporate ensemble coding mechanisms within a dictionary learning approach via coherence and structural constraints to learn meaningful crowd prototypes; (3) we evaluate the proposed approach on multiple crowd perception problems from collective or abnormal crowd detection to tracking individuals in crowded scenes. Experimental results on crowd datasets demonstrate competitive results on par or better than state-of-the-art approaches.

2 The Approach

2.1 Hierarchical Model of Crowd Processing

An overview of the system is shown on Fig. 1. The basic visual representation is based on [28], which we only review briefly here. The model starts with motion-sensitive simple (S1) and complex (C1) units similar to those found in the primary visual cortex. In [28], Jhuang et al. compared several implementations of motion-sensitive S1 units. Here, we consider their implementation based on optical flow, because it is particularly amenable to extending existing approaches for crowd perception [2,5,15]. Specifically, we build a population of motion-sensitive simple (S1) units tuned to both speed and motion direction using the optical flow estimated from local space-time 3D volumes. Depending on the application (see later), these volumes are sampled either at random locations or at locations returned by the gKLT tracker as done in [11].

Let $\theta_{i,j}$ and $v_{i,j}$ denote the direction and velocity of the optical flow at image location (i, j). As done in [27,28], simple (S1) unit responses are then obtained using the following quantization:

$$r_{S1}^{i,j}(\theta_p, v_p) = \left\{ \frac{1}{2}[1 + \cos(\theta_{i,j} - \theta_p)] \right\}^q \times \exp(-|v_{i,j} - v_p|), \tag{1}$$

where $\theta_p \in \{0°, 90°, 180°, 270°\}$ and $v_p \in \{3, 6\}$ correspond to the preferred direction and speed of the unit, and the constant q controls the width of the tuning curve (here $q = 2$, see [27] for details). In the following stage, C1 unit responses are computed via a local max pooling on the S1 unit responses across both speeds and a local $l \times l$ spatial neighborhood.

In subsequent processing stages, units of higher visual complexity emerge after an additional *template-matching* (S2 units) as well as an *invariance-pooling* (C2 units) stage, increasing both the selectivity and invariance properties of the underlying model units. The response of S2 units is obtained by convolving C1 maps across all motion directions with a dictionary of stored prototypes. Originally, the dictionary of K S2 prototypes is learned via a simple random sampling procedure. Here, instead, we propose to learn crowd prototypes via sparse coding methods, which we describe next.

2.2 Learning Crowd Prototypes

Given a set of N input vectors $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N] \in \mathbb{R}^{D \times N}$, learning a sparse dictionary of coding elements can be formulated as the following optimization problem:

$$\mathbf{B}^*, \mathbf{S}^* = \arg\min_{\mathbf{B},\mathbf{S}} \|\mathbf{R} - \mathbf{B}\mathbf{S}\|_2^2 + \lambda \sum_i \|\mathbf{s}_i\|_1, \text{s.t.} \quad \forall i, \mathbf{s}_i \succeq 0, \tag{2}$$

where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K] \in \mathbb{R}^{D \times K}$ is a matrix that contains the learned basis functions as column vectors and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N] \in \mathbb{R}^{K \times N}$ is a matrix containing the corresponding linear coefficients. λ is a constant to control the tradeoff between the reconstruction error and the sparsity of the underlying representation.

We propose to incorporate the idea of ensemble coding in the form of two additional penalty terms embedded in Eq. 2. Cognitive psychology experiments have suggested the existence of pre-attentive pooling mechanisms used by our visual system to force chaotically moving crowds to cohere into a unified and visually appealing Gestalt. Psychophysics experiments have shown that participants rapidly pool information from multiple walkers to estimate the heading of a crowd [25]. Here we model this phenomenon via a *structural neighborhood cohesion* term, which forces input patterns to converge towards a similar interpretation and a *neighborhood manifold coherence* term, which incorporates explicit pooling mechanisms over output vectors to yield a locally more stable code.

These two constraints are embedded in the following optimization problem:

$$\mathbf{R}^*, \mathbf{B}^*, \mathbf{S}^* = \arg\min_{\mathbf{R},\mathbf{B},\mathbf{S}} \underbrace{\|\mathbf{R} - \mathbf{BS}\|_F^2}_{recon.\ error} + \underbrace{\lambda \sum_{i=1}^{N} \|\mathbf{s}_i\|_1}_{sparsity\ term} + \underbrace{\beta \sum_{i=1}^{N} \|\mathbf{r}_i - \mathbf{r}'_i\|_M^2}_{structural\ term}$$
$$+ \underbrace{\gamma \sum_{i=1}^{N} \sum_{j=1}^{N} \|\mathbf{s}_i - \mathbf{s}_j\|^2 \mathbf{W}_{ij}}_{coherence\ term}, \text{ s.t. } \|\mathbf{b}_k\|^2 \leq c, k = 1, \ldots, K. \quad (3)$$

Here $\mathbf{r}'_i$ corresponds to the average over all $\mathbf{r}_i$ within the spatial neighborhood of unit i. λ, β and γ are constants used to trade the weights between the various regularization terms. The learning algorithm is initialized by setting up vectors of model C1 unit responses (tuned to different directions of motion over a local spatial neighborhood) as $\mathbf{r}_j$, such that $\mathbf{R}$ is the matrix containing all N C1 unit vectors as columns. $\mathbf{S}$ is the response of the S2 units.

In the above objective function, the first term is an estimate of the reconstruction error when encoding the S2 unit responses using the learned prototypes and associated coefficients. The second term corresponds to a standard sparsity constraint on the coefficients, which constrains the number of prototypes actually used to encode a given visual sample $\mathbf{r}_i$ to be small. We formulate the coherence constraint as a graph-based Laplacian regularization problem [32] while we formulate the structural constraint as a generalized Tikhonov regularization problem [19]. The coherence constraint should, in principle, help build a visual representation that takes into account the local manifold structure of the data enforcing local consistency of the flow. The structural term should help learn crowd patterns with locally similar trajectories from individuals. This can be also thought of as a denoising term effectively smoothing out the local motion flow towards a common vector $\mathbf{r}'_i \in \mathbf{R}'$ over a local spatial neighborhood (weighted by a Gaussian function over space):

$$\mathbf{r}'_i = \arg\min_{\mathbf{r}_i} \frac{1}{d} \sum_{j \in \mathcal{N}(i)} \exp(-\frac{\|\mathbf{r}_i - \mathbf{r}_j\|^2}{2\sigma^2}), \quad (4)$$

where $\mathcal{N}(i)$ denotes the set of indexes for the d nearest neighbors around $\mathbf{r}_i$ and σ is a constant.

Because the objective function in Eq. 3 is not convex with respect to $\mathbf{R}$, $\mathbf{B}$ and $\mathbf{S}$, we use a two-alternative minimization approach, alternatively optimizing one variable while fixing the others (see Algorithm 1). In step (A), the matrix of C1 response vectors $\mathbf{R}$ are computed after fixing $\mathbf{B}$ and $\mathbf{S}$. Eq. 3 can be rewritten by replacing the fixed term $\mathbf{BS}$ with $\mathbf{b}$ as detailed in the following matrix form:

$$\mathbf{R}^* = \arg\min_{\mathbf{R}} \|\mathbf{R} - \mathbf{b}\|_F^2 + \beta \|\mathbf{R} - \mathbf{R}'\|_M^2, \quad (5)$$

where $\|\mathbf{R} - \mathbf{R}'\|_M^2 = (\mathbf{R} - \mathbf{R}')^T \mathbf{Q}^{-1} (\mathbf{R} - \mathbf{R}')$ is the Mahalanobis distance between $\mathbf{R}$ and $\mathbf{R}'$. $\mathbf{Q}$ is the covariance matrix computed over $\mathbf{R}$.

Algorithm 1. Crowd prototype learning

input: Given N C1 unit reponse vectors $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N] \in \mathbb{R}^{D \times N}$ and fixed parameters;
Initialize $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K] \in \mathbb{R}^{D \times K}$ and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N] \in \mathbb{R}^{K \times N}$;
repeat
  Step (A):
  Given $\mathbf{B}$, $\mathbf{S}$, compute $\mathbf{R}'$ by pooling over the d nearest neighbors of $\mathbf{R}$ according to Eq. 4;
  Solve $\mathbf{R}^*$ via generalized Tikhonov regularization (Eq. 5);
  Update $\mathbf{R}$ with $\mathbf{R}^*$ (Eq. 6),;
  Step (B):
  Given $\mathbf{R}$, solve for $\mathbf{B}$, $\mathbf{S}$ (Eq. 8) using the *feature sign* search algorithm [33];
  Update $\mathbf{B}$, $\mathbf{S}$;
  Iteration number i++;
until *Change in* $\mathbf{S}$ *between 2 successive iterations is smaller than* ε *or max iteration number reached*;
output: Optimized $\mathbf{R} \in \mathbb{R}^{D \times N}$, crowd prototypes $\mathbf{B} \in \mathbb{R}^{D \times K}$, S2 response coefficients $\mathbf{S} \in \mathbb{R}^{K \times N}$;

In Step (A), after the neighborhood pooling step (see Eq. 4) at each iteration, a closed-form solution can be computed using the generalized Tikhonov regularization as:

$$\mathbf{R}^* = \mathbf{R}' + (\mathbf{I} + \beta \mathbf{Q}^{-1})^{-1}(\mathbf{b} - \mathbf{R}'). \tag{6}$$

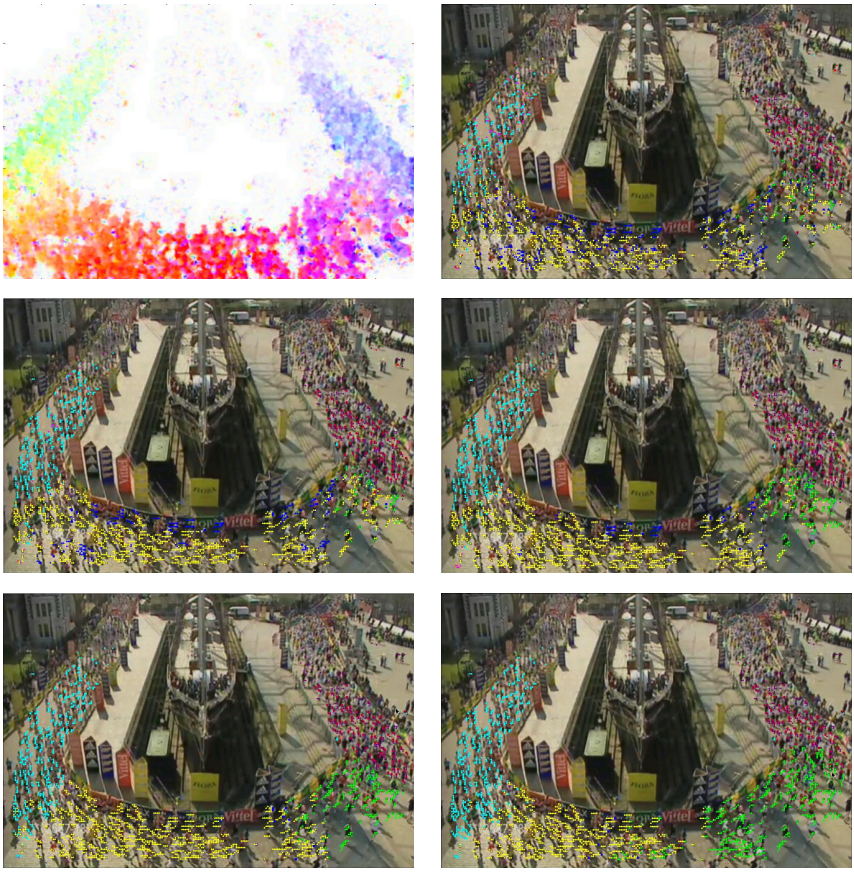
In Step (B), we follow the approach described in [32]. Let $\mathbf{W} \in \mathbb{R}^{N \times N}$ be a nearest neighbor indicator matrix ($\mathbf{W}_{ij} = 1$ if $\mathbf{r}_i$ and $\mathbf{r}_j$ are nearest neighbors and $\mathbf{W}_{ij} = 0$ otherwise). The degree of $\mathbf{r}_i$ is defined as $d_i = \sum_{j=1}^{N} \mathbf{W}_{ij}$ and $\mathbf{D} = diag(d_1, \ldots, d_N)$. This term can be rewritten as follow:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \|\mathbf{s}_i - \mathbf{s}_j\|^2 \mathbf{W}_{ij} = Tr(\mathbf{S}^T \mathbf{L} \mathbf{S}), \tag{7}$$

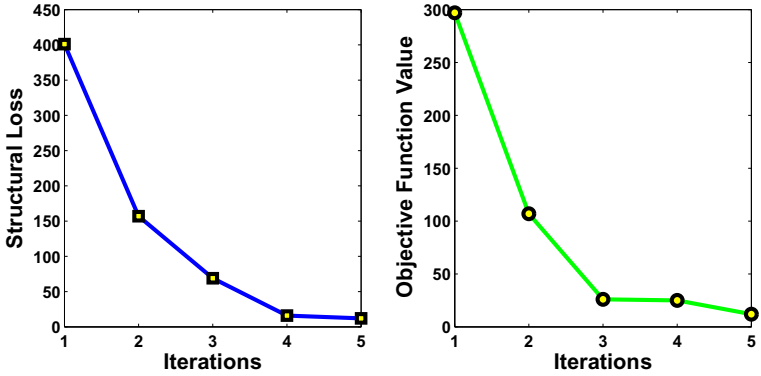
where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. By fixing $\mathbf{R}$ and incorporating the Laplacian regularizer, $\mathbf{B}$ and $\mathbf{S}$ can be updated according to:

$$\arg\min_{\mathbf{B},\mathbf{S}} \|\mathbf{R} - \mathbf{B}\mathbf{S}\|_F^2 + \lambda \sum_i \|\mathbf{s}_i\|_1 + \gamma Tr(\mathbf{S}^T \mathbf{L} \mathbf{S}),$$
$$\text{s.t.} \|\mathbf{b}_k\|^2 \leq c, k = 1, \ldots, K. \tag{8}$$

The above optimization is a typical laplacian regularization problem, which can be solved using the *feature sign* search algorithm [33]. As $\beta \to 0$, Eq. 3 degenerates into a typical graph-based sparse coding approach. Similarly as $\beta, \gamma \to 0$, Eq. 3 degenerates to a standard sparse coding optimization. In general, we have found the optimization procedure to converge quickly within 5 iterations (see Fig. 2 for a representative example). Thus, in all subsequent experiments we fixed the maximum number of iterations to $n = 5$.



(a) Sample convergence results on the Marathon sequence. Leftmost frame: standard optical flow followed by the prototype assignments for the first 5 iterations.



(b) Corresponding structural loss and value of the objective objective function.

Fig. 2. Illustrative convergence results for the proposed approach

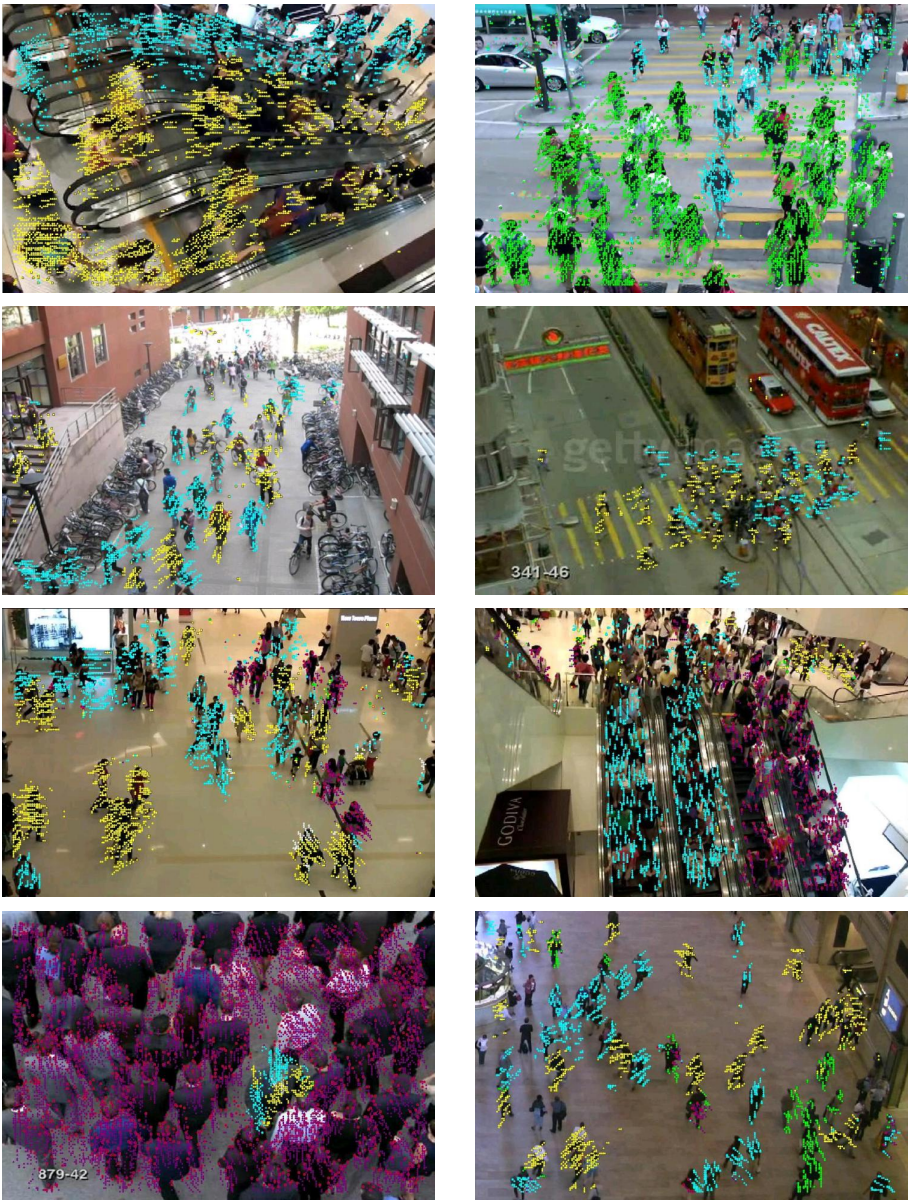


Fig. 3. Representative examples of the learned prototypes. Shown are sample frames from the Collective Motion Dataset [11] overlaid with color coded symbols (best seen in color) indicating the closest prototype for the corresponding location.

Given a set of C1 unit responses $\mathbf{R}$ and a dictionary of prototypes $\mathbf{B}$, one can compute the corresponding reconstruction coefficients $\mathbf{S}$ as S2 responses. Fig. 3 shows sample frames overlaid with symbols which indicate the prototype associated with the largest coefficient for that location. From visual inspection, it is clear that the learned prototypes are able to selectively capture a variety of crowd behaviors including crossing, lane forming, etc.

A sparse, invariant representation for crowd patterns can be obtained by computing the maximum coefficient over a spatial (and possibly temporal) neighborhood for each prototype at the next (C2) stage. In order to evaluate the effectiveness of the proposed approach, we carry out three experiments including: (1) the detection of collective motion patterns, (2) the detection of abnormal crowd behaviors and (3) the tracking of individuals in crowds. Additional results which could not be included because of space constraints can be found online at `http:/serre-lab.clps.brown.edu/resource/zhangetalhbu2014`.

3 Experiments

3.1 Abnormal Event Detection

Video dataset: We consider the UMN dataset (`http:/mha.cs.umn.edu/Movies/Crowd-Activity-All.avi`), which contains 11 video clips containing crowded escape video events acquired in 3 different scenarios. Each video begins with a normal behavior and ends with a panic escape. We resized all video frames to 120×160 pixels for computational efficiency.

Detection: For learning crowd prototypes, we sampled C1 unit vectors of size $5 \times 5 \times 4$ from the first 10 frames in each video sequence. Rather than considering random locations as in [28], we here sampled at keypoints returned by the gKLT tracker as done in [11] for crowds. We further pruned out vectors corresponding to locations with little activity by discarding vectors with a norm below a fixed threshold ($\theta = 0.1$).

We trained a dictionary of prototypes with $K = 30$. After the S2 stage, the maximum coefficient for each prototype over all locations and all frames were computed to yield C2 units, which can then be used as a compact visual representation for crowds. For classification, we used an SVM with an RBF kernel using the standard training/test data split to classify events as normal vs. abnormal and the accuracy measure described in [21].

Evaluating the different penalty terms: To assess the benefit of the different penalty terms in the proposed optimization function, we first sample a fixed set of C1 unit vectors to be used for all sparse coding based approaches. Fig. 4 shows a comparison of the system accuracy using different regularization terms: A basic Sparse Coding (SC) as described in [33] without either the coherence or structural constraint, two implementations of the proposed algorithm with the coherence term only (i.e., without the structural term) based on Laplacian Sparse Coding (LapSC) [34] and Graph-based Sparse Coding(GraphSC) [32] for varying

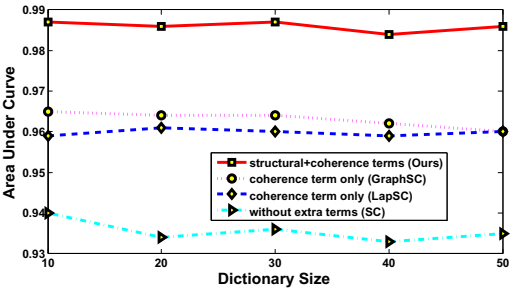


Fig. 4. Evaluation of the different penalty terms used in the proposed optimization function

Table 1. AUC measures for the detection of abnormal behavior on the UMN dataset

Method	**Our approach**	Cong et al. (SRC) [21]	Cui et al. (IEP) [8]	Mehran et al. 2009 (SF) [3]	Mehran et al. 2010 (SP) [9]	Optical Flow
AUC	**0.987**	0.99	0.985	0.96	0.90	0.86

dictionary sizes. It is pretty clear from this experiment that both constraints are indeed useful and that the proposed algorithm significantly outperform a vanilla sparse coding model.

Comparison with state-of-the-art approaches: Table 1 shows the accuracy of the proposed approach measured by the area under the ROC (AUC) together with a comparison with state-of-the-art approaches on the UMN dataset. Accuracy measures for the benchmark systems were those reported in the original studies [8,3,9,21]. The proposed approach achieve results that are on par or better than state-of-the-art systems including the Interaction Energy Potentials [8] (IEP), Social Force [3] (SF), Streakline Potential [9] (SP) and a standard Optical Flow (OF) based approach. The accuracy of our approach is only slightly lower than the Sparse Reconstruct Cost (SRC) method [21] despite the fact that the SRC uses MHOF as an input, which is much more robust to illumination, distortion and noise compared to the optical flow used here. Future work should compare the two approaches using the same exact inputs.

3.2 Collectiveness Classification

Video dataset: Here we consider the Collective Motion Dataset [11], which contains 413 videos from 62 crowded scenes including malls, traffic scenes, escalators, campuses, etc. Each video sequence contains 100 frames with ground truth annotations corresponding to 3 different levels of collectiveness — low, medium and high — obtained from 10 human observers. We used the same procedure as in [11] where a classifier is trained to discriminate between high vs. low, high vs. medium and medium vs. low.

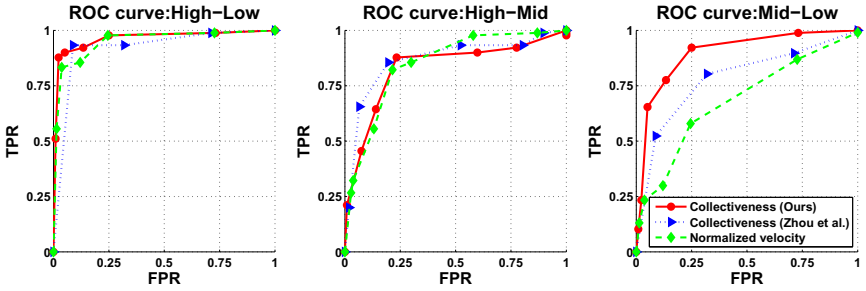


Fig. 5. ROC curves for the classification of collectiveness levels. We compare a "prototype" score P derived using the proposed approach with a"collectiveness" score C and the "normalized velocity" V (see text for details).

Collectiveness score: For learning crowd prototypes, we sampled C1 unit vectors of size 8×8 ($\times 4$ directions of motion) as done in [28] from the entire database. Here, however, we sampled at keypoints returned by the gKLT tracker as done in [11] for crowds as opposed to randomly sampled locations as done in [28]. We further pruned out vectors corresponding to locations with little activity, i.e., vectors with a norm below a fixed threshold ($\theta = 0.1$) were discarded.

After the S2 stage, units are then pooled over a 8×8 spatial neighborhood to yield C2 units, which can then be used as a compact visual representation for crowds. We first consider an application to assessing the collectiveness of a crowd [11]. We used the C2 unit responses to compute a collectiveness metric score P as proposed by Zhou et al. [11]:

$$P = \frac{1}{|\Omega|} \sum \mathbf{e}^T ((\mathbf{I} - z\mathbf{W}_p)^{-1} - \mathbf{I})\mathbf{e}, \tag{9}$$

where $\mathbf{W}_p$ corresponds to the adjacency matrix of the graph obtained by computing the χ^2 distance between C2 unit responses i and j over the set Ω. $\mathbf{e}$ is a vector with all elements set to 1.

Evaluation: Fig. 5 shows a comparison between the collectiveness score computed using the proposed representation (P) with two collectiveness score C and the normalized velocity V described in [11]. ROC curves are presented for 3 levels of collectiveness as done in [11]: Low, Medium and High collectiveness. For the High-Medium and High-Low categories, our prototypes perform on par with the state-of-the-art. This may reflect the ability of the proposed visual representation to distinguish different levels of dynamic motion, while preserving the consistency and structure of the crowd. Furthermore, our approach outperforms other approaches on the more challenging Medium-Low category.

3.3 Tracking in Crowded Scenes

Tracking framework: Because of the high similarity between targets and distractors, as well as the presence of significant occlusions, tracking in crowds is a

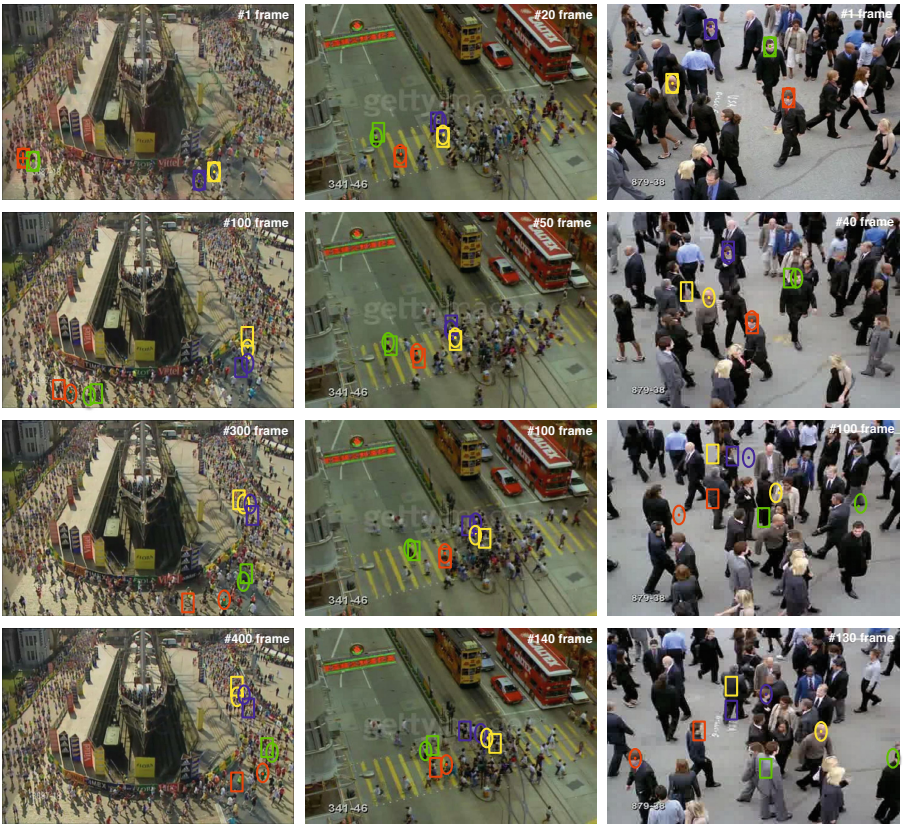
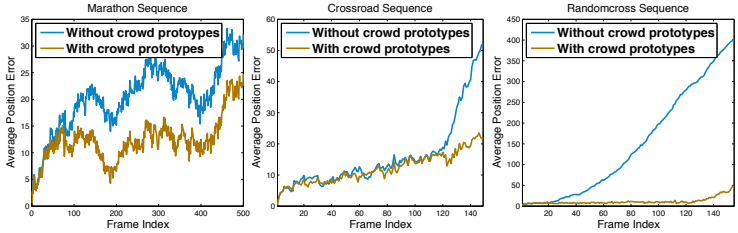


Fig. 6. Tracking results on 3 sequences for comparison between the proposed approach (circles) vs. the approach by Zhang et al [35] (squares). The ground truth is shown with dots. Tracking results for different subjects are marked with different colors.

very challenging problem. Classical single- and multi-target tracking approaches [35,5,15] have focused on extracting discriminative appearance models, often overlooking the problem of modeling the target's individual movement. These methods are usually based on a simple dynamic model with a smooth motion prior and additive Gaussian noise to predict the location of the target:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{v}_t + \mathbf{n} \tag{10}$$

where $\mathbf{x}_t$ corresponds to the target location, $\mathbf{v}_t$ the target 2D motion vector and $\mathbf{n}$ is Gaussian noise. Typically, $\mathbf{v}_t$ is computed using the state of the individual target from previous times, e.g., $\mathbf{v}_t = \mathbf{x}_{t-1} - \mathbf{x}_{t-2}$. Because of random jitter in the predicted target location, the computed motion vector is usually relatively noisy. Here, instead, we propose to compute the motion vector based on the average motion vector computed for all keypoints within the sampling region associated with the prototype with the highest assignment count. These prototypes are



Method	Marathon	Crossroad	Randomcross
Zhang et al. [35]	23.8	33.6	128.4
Kratz et al. [5]	15.6	**3.56**	17.3
Rodriguez et al. [15]	47.8	25.9	29.9
Ours	**8.89**	5.43	**10.7**

Fig. 7. Top: Average position error curves for the proposed sparse coding approach and comparison with baseline. Bottom: Comparison between tracking approaches using the average position error computed over entire sequences.

learned by sampling C1 unit vectors for 5 consecutive frames with a dictionary of size $K = 8$. These samples are then updated over time and prototypes are learned anew for every frame. We implement a multi-target tracking algorithm for crowded scenes by extending the real-time tracker described in [35] as a base system. We extend this single-target tracker (originally based on a brute-force search approach) with a particle filtering framework.

Evaluation: We assess the accuracy of the proposed tracking approach and compare it to a baseline system, which also uses particle filtering, but with a simple state transition model (Eq. 10) as well as several state-of-the-art systems [5,15] for crowd tracking. Our evaluation is based on the crowded sequences used in [2,3]. Accuracy is measured by the Average Position Error (APE), which corresponds to the average difference (in pixels) between the position of the tracked object and the corresponding ground truth. Qualitative results are shown in Fig. 6 together with the average position errors for all methods in Fig. 7.

4 Conclusion

In this paper, we have extended a biological model of motion processing [28] from the recognition of individual human activity to group behaviors and described a new method to learn a dictionary of crowd prototypes. Motivated by human studies on crowd perception, our approach incorporates ensemble coding principles via structural and local coherence constraints within a sparse coding model. We have demonstrated the wide applicability of the approach to several problems in crowd perception. Experiments on public datasets demonstrate that the proposed model exhibits competitive performance against state-of-the-art approaches.

Acknowledgments. This work was supported by ONR grant (N000141110743) and NSF early career award (IIS-1252951) to TS. Additional support was provided by the Robert J. and Nancy D. Carney Fund for Scientific Innovation and the Center for Computation and Visualization (CCV) at Brown University. YZ and QH were supported in part by the National Basic Research Program of China (973 Program, 2012CB316400) and the National Natural Science Foundation of China (61025011, 61133003, 61300111, 61332016 and 61035001). YZ was funded by the China Scholarship Council.

References

1. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical Review E 51(5), 4282 (1995)
2. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: CVPR (2007)
3. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: CVPR (2009)
4. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.J.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
5. Kratz, L., Nishino, K.: Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In: CVPR (2010)
6. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: CVPR (2010)
7. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who are you with and where are you going? In: CVPR (2011)
8. Cui, X., Liu, Q., Gao, M., Metaxas, D.: Abnormal detection using interaction energy potentials. In: CVPR (2011)
9. Mehran, R., Moore, B.E., Shah, M.: A streakline representation of flow in crowded scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 439–452. Springer, Heidelberg (2010)
10. Kratz, L., Nishino, K.: Going with the flow: Pedestrian efficiency in crowded scenes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 558–572. Springer, Heidelberg (2012)
11. Zhou, B., Tang, X., Wang, X.: Measuring crowd collectiveness. In: CVPR (2013)
12. Wu, S., Moore, B.E., Shah, M.: Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In: CVPR (2010)
13. Solmaz, B., Moore, B., Shah, M.: Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE TPAMI 34(10), 2064–2070 (2012)
14. Hospedales, T., Gong, S., Xiang, T.: Video behaviour mining using a dynamic topic model. International Journal of Computer Vision 98(3), 303–323 (2012)
15. Rodriguez, M., Ali, S., Kanade, T.: Tracking in unstructured crowded scenes. In: ICCV (2009)
16. Lin, D., Grimson, E., Fisher, J.: Learning visual flows: A lie algebraic approach. In: CVPR (2009)
17. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In: CVPR (2009)
18. Andrade, E., Blunsden, S., Fisher, R.: Hidden markov models for optical flow analysis in crowds. In: ICPR (2006)

19. Zhao, X., Gong, D., Medioni, G.: Tracking using motion patterns for very crowded scenes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 315–328. Springer, Heidelberg (2012)
20. Zen, G., Ricci, E.: Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes. In: CVPR (2011)
21. Cong, Y., Yuan, J., Liu, J.: Sparse reconstruction cost for abnormal event detection. In: CVPR (2011)
22. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: ICCV (2013)
23. Zhao, B., Fei-Fei, L., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In: CVPR (2011)
24. Zen, G., Ricci, E., Sebe, N.: Exploiting sparse representations for robust analysis of noisy complex video scenes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 199–213. Springer, Heidelberg (2012)
25. Sweeny, T.D., Haroz, S., Whitney, D.: Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. Journal of Experimental Psychology: Human Perception and Performance 39(2), 329 (2013)
26. Crouzet, S.M., Serre, T.: What are the visual features underlying rapid object recognition? Frontiers in Psychology 2 (2011)
27. Giese, M.A., Poggio, T.: Neural mechanisms for the recognition of biological movements. Nature Reviews Neuroscience 4(3), 179–192 (2003)
28. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV (2007)
29. Taylor, G.W., Fergus, R., LeCun, Y., Bregler, C.: Convolutional learning of spatio-temporal features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 140–153. Springer, Heidelberg (2010)
30. Zhang, Y., Qin, L., Yao, H., Huang, Q.: Abnormal crowd behavior detection based on social attribute-aware force model. In: ICIP (2012)
31. Zhang, Y., Qin, L., Yao, H., Xu, P., Huang, Q.: Beyond particle flow: Bag of trajectory graphs for dense crowd event recognition. In: ICIP (2013)
32. Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., Cai, D.: Graph regularized sparse coding for image representation. IEEE TIP (2011)
33. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: NIPS (2006)
34. Gao, S., Tsang, I.W., Chia, L.T., Zhao, P.: Local features are not lonely–laplacian sparse coding for image classification. In: CVPR (2010)
35. Zhang, K., Zhang, L., Yang, M.-H.: Real-time compressive tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)