COMPLEX SCENE ANALYSIS USING HIERARCHICAL SPARSE CONCEPT REPRESENTATION

Sanghyuk Park[1], Jaesik Yoon[2], Chang D. Yoo[3] , and Jaecheol Kwon[4]

[1,2,3] Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[4] Future Research Lab., Korea Telecom R&D Center, Seoul, Republic of Korea
[1]shine0624@kaist.ac.kr, [2]jaesik817@kaist.ac.kr, [3]cdyoo@ee.kaist.ac.kr, [4]jckwon@kt.com

This paper considers a hierarchical sparse concept representation for complex scene analysis. Low-level visual features, foreground pixels and optical flows, are commonly used for the scene analysis. However, these features are sensitive to noise and have high-dimension to represent complex behavior of various objects. The considered hierarchical sparse concept representation aims to cope with noise of low-level features and reducing uncertainty in determining behavior patterns. In this paper, the task of complex scene analysis is formalized as discovering high-level behavior patterns which represent video context. The experimental results show that the considered algorithm yields good performance using complex video datasets.

Introduction

Automatic detecting representative behavior pattern has become a challenging task for complex scene analysis (CSA). In the past few years, various algorithms have been proposed to extract semantic behavior patterns from video scenes. They focused on analyzing complex and crowded video scene based on discovered specific patterns from low-level visual features. Previous algorithms usually consider video data as a set of visual features which are represented using a histogram of occurrences of feature, and the extracted features are directly used for CSA. However, these feature based algorithms have mainly two problems: (1) low-level visual features have limitation to represent high-level semantic information of behavior patterns. Hence, high dimensional feature space is required and computational complexity is increased; and (2) low-level visual features are sensitive to noise and some occlusion from movements of various objects in the complex video scene.

Recently, statistical machine learning based algorithms have shown good performance in the task of CSA based on: multi-scale methods [1, 2], hierarchical methods [3, 4, 5], Bayesian methods [6, 7], etc. These algorithms try to implicitly handle varying noise and uncertainty in discovering semantic behavior patterns from high dimensional visual feature spaces.

This paper considers a hierarchical sparse concept representation (HSCR) algorithm for the complex video scene analysis. In this paper, the task of CSA is considered as the problem of discovering a set of high-level semantic behavior patterns using hierarchical dictionary learning and sparse concept representations. The considered HSCR algorithm aims to cope with noise of high dimensional low-level visu-al features and reducing uncertainty in discov-ering semantic behavior patterns while pre-serving discriminative information of primary patterns and dimensionality reduction.

Non-object-based bag-of-words representation is used as low-level visual features similar to previous literature. However, the considered HSCR algorithm tries to discover multi-level behavior patterns instead of using noisy low-level visual features directly. Complex video scenes usually contain multiple movements of various objects over time and space, this paper assumes that these movements can be categorized into low-level and high-level behavior patterns. Low-level behaviors (e.g. moving of vehicle and pedestrians) can be characterized using representative movements which occur in a small range of spatial and temporal regions. High-level behaviors (e.g. traffic flows by signal) can be described as periodically repeated patterns using a set of low-level behavior patterns. Considering these hierarchical structures of behavior patterns can be helpful to analyze video scene more accurately than using single behavior patterns. In considered HSCR, behavior patterns of each level can be described using small number of basis which related to most representative behavior pattern in each level. For this, hierarchical basis learning and sparsity constraint are considered not only reducing dimensionality of features but also preserving discriminative information.

Review of Sparse Representation

Let $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_N]\in\Re^{M\times N}$ is a set of *N* data with high-dimensional feature space. Matrix factorization (MF) algorithms [8, 9, 10] are common approaches to condense data by discovering a set of new basis vector and the new representation with respect to the new basis for each data. The aim of MF is finding two matrices, dictionary $\mathbf{D}$ and sparse representation $\mathbf{A}$, whose product can well approximate $\mathbf{X}$. Given a dictionary $\mathbf{D}=[\mathbf{d}_1,\ldots,\mathbf{d}_K]\in\Re^{M\times K}$, the sparse representation $\mathbf{A}=[\mathbf{a}_1,\ldots,\mathbf{a}_N]\in\Re^{K\times N}$ for $\mathbf{X}$ can be obtained by solving:

$$\mathbf{A}=\operatorname{argmin}_{\mathbf{A}}\|\mathbf{X}\text{-}\mathbf{DA}\|_{\mathrm{F}}^2 \text{ s.t.}\forall i,\|\mathbf{a}_i\|_0\leq\varepsilon \quad (1)$$

Where, $\|\mathbf{X}\text{-}\mathbf{DA}\|_{\mathrm{F}}^2$ denotes the reconstruction error, and $\|\mathbf{a}_i\|_0\leq\varepsilon$ is the sparsity constraint. Each column of $\mathbf{D}$ is a basis vector and each column of $\mathbf{A}$ is the K dimension representation of the original input data with respect to the new basis $\mathbf{D}$.In this sense, MF can be regarded as a dimensionality reduction algorithm since it tries to reduce the dimension from *M* to *K*. The performance of sparse representation $\mathbf{A}$ depends critically on the constructed $\mathbf{D}$.

Hierarchical Sparse Concept Representation

Given a data set of high dimensional feature $\mathbf{X}$, the purpose of HSCR algorithm is finding low dimensional high-level behavior patterns while reducing noise and dimensionality of features with reducing information loss. It can be obtained by solving the minimization problem as follows:

$$\min_{\mathbf{D}_{\mathrm{L}},\mathbf{D}_{\mathrm{H}},\mathbf{S}}\sum_{i=1}^{N}[\|\mathbf{x}_i-\mathbf{D}_{\mathrm{L}}\mathbf{D}_{\mathrm{H}}\mathbf{s}_i\|_2^2+\phi\|\mathbf{D}_{\mathrm{H}}\mathbf{s}_i\|_1] \quad (2)$$

where the dictionary $\mathbf{D}_{\mathrm{L}}\in\Re^{M\times Z}$ is the basis of low-level behavior patterns and $\mathbf{D}_{\mathrm{H}}\in\Re^{Z\times K}$ is the basis of high-level behavior patterns. The matrix $\mathbf{S}\in\Re^{K\times N}$ indicates HSCR of input data $\mathbf{X}$ and ϕ is the regularization parameter. Eq. (2) is not jointly convex to $\mathbf{D}_{\mathrm{L}},\mathbf{D}_{\mathrm{H}},\mathbf{S}$. However, it is convex with respect to each of them if others are fixed. Hence, the HSCR can be designed using an iterative algorithm to alter-natively optimize each matrix base.

The HSCR algorithm has three-step in terms of dimensionality reduction and MF based on sparse concept coding algorithm [12]. The first step of HSCR is low-level behavior basis learning by exploring the low-dimensional intrinsic geometric structure of the input data instead of using visual feature space directly. Spectral regression [11, 12] is used for reducing dimension of data and finding manifold of the ambient space to model the local geometric structure of data with weight $\mathbf{W}$. Similarly, in the previous MF algorithm, considered HSCR tries to learn low-level behavior basis matrix $\mathbf{D}_{\mathrm{L}}$ which can well fit to $\mathbf{Y}=\{\mathbf{y}_i\}_{i=1}^{Z}\in\Re^{N\times Z}$. It can be obtained by solving the optimization problem as follows:

$$\min_{\mathbf{D}_{\mathrm{L}}}\|\mathbf{Y}-\mathbf{X}^T\mathbf{D}_{\mathrm{L}}\|_{\mathrm{F}}^2+\alpha\|\mathbf{D}_{\mathrm{L}}\|_2^2 \quad (3)$$

where $\|\mathbf{D}_{\mathrm{L}}\|_2^2$ is a regularization term to avoid over-fitting and α is the regularization parameter. Where, the optimal $\mathbf{Y}$ can be obtained by solving Eq.(4) using the minimum eigenvalue eigen-problem of $\mathbf{LY}=\lambda\mathbf{D}_{\mathrm{L}}\mathbf{Y}$.

$$\mathbf{Y}^*=\operatorname{argmin}(\mathbf{Y}^T\mathbf{LY})/(\mathbf{Y}^T\mathbf{D}_{\mathrm{L}}\mathbf{Y}) \quad (4)$$

Here, $\mathbf{L}=\mathbf{D}_{\mathrm{L}}-\mathbf{W}$and $\mathbf{D}_{ii}=\sum_j\mathbf{W}_{ij}$.The weight matrix $\mathbf{W}$ is constructed using EMD distance. After obtaining low-level behavior basis $\mathbf{D}_{\mathrm{L}}$, the second step is computing the low-level sparse concept representation $\mathbf{S}_{\mathrm{L}}$ by solving the optimization problem as follows:

$$\min_{\mathbf{S}_{\mathrm{L}}}\sum_{i=1}^{N}[\|\mathbf{x}_i-\mathbf{D}_{\mathrm{L}}\mathbf{s}_{\mathrm{L,i}}\|_2^2+\beta\|\mathbf{s}_{\mathrm{L,i}}\|_1] \quad (5)$$

where $\mathbf{S}_{\mathrm{L}}\approx\mathbf{D}_{\mathrm{H}}\mathbf{S}$ and $\|\mathbf{s}_{\mathrm{L,i}}\|_1$ is the L1-norm regularization to enforce the sparsity. The Least Angel Regression [13] is used for solving the optimization problem in Eq. (5). In this step, $\mathbf{S}_{\mathrm{L}}$ can be regarded as representative low-level behavior patterns using small number of basis of $\mathbf{D}_{\mathrm{L}}$. The last step of HSCR algorithms is learning of high-level behavior basis and finding HSCR by solving the optimization problem as follows:

$$\min_{\mathbf{D}_{H},\mathbf{S}_{H}} \sum_{i=1}^{N}[\|\mathbf{s}_{L,i}-\mathbf{D}_{H}\mathbf{s}_{H,i}\|_{2}^{2}+\gamma\|\mathbf{s}_{H,i}\|_{1}] \quad (6)$$

where $\mathbf{D}_{H}$ indicates the basis matrix of high-level behaviors and $\mathbf{S}_{H}$ is the HSCR using $\mathbf{D}_{H}$. For this, $\mathbf{D}_{H}$ can be obtained using Eq.(3) with respect to $\mathbf{S}_{L}$, and $\mathbf{S}_{H}$is calculated using Eq.(5) with respect to $\mathbf{S}_{L}$ and $\mathbf{D}_{H}$, iteratively. Finally, the considered HSCR algorithm can be convert original high-dimensional low-level features $\mathbf{X}$ to low-dimensional HSCR $\mathbf{S}_{H}$, while removing noise using dimensionality reduction and preserving discriminative information.

Experiments and Results

In this section, we investigate the effectiveness of the considered HSCR algorithm for CSA. The task of CSA is considered as a multi -class clustering problem similar to previous literatures [2, 3, 7, 14, 15]. For this, we convert each high-dimensional test data into low-dimensional HSCR by using the learned low-level and high-level behavior basis from training data. Then, each converted test data is clustered into the *K* number of primary high-level behavior patterns by the nearest centroid criterion. To verify the effectiveness of HSCR, various experiments were conducted on the publicly available QMUL [3] and CVBASE '06 [16] dataset. All dataset contains complex behavior patterns of a large number of objects such as vehicles and human movements. A ***detailed*** description of experimental parameter is summarized in Table 1.

Table 1. Parameter setup (train)

Dataset	M	N	Z	K	α	β	γ
Junction-1	4176	73	8	2	0.1	0.02	0.4
Roundabout	4176	146	8	2	0.2	0.02	0.56
Basketball	4292	50	4	2	0.2	0.16	0.08
Handball	4524	100	8	2	0.2	0.32	0.1

The Traffic datasets: As shown in Figure 1, The QMUL dataset contains complex traffic scenes and have been extensively used in previous CSA literatures [2, 3, 7, 14, 15]. Each dataset was recorded by a stationary camera with a resolution of 360×288 (25 FPS).

(a) Junction-1 (b) Roundabout

Fig. 1. QMUL datasets

Each video clip was spatially quantized into 36×29 cells and 4 directions of quantized optical flow were extracted in each cell. Hence, the size of low-level feature dimension M is 4176. The test dataset consist of 39, 59 video clips for the Junction-1 and Roundabout, res-pectively. For fair comparison between the considered HSCR algorithm and the state-of-the-art algorithms, the same datasets and ground truth labels [3, 14, 15] are used. For the database in hand, there are two main temporal phases of the primary behavior patterns: vertical and horizontal traffic flows. The quantitative results of the clustering accuracy based on primary high-level behavior patterns are represented in Table 2.

Table 2. Clustering accuracy (QMUL)

Algorithms	Junction-1 (%)	Roundabout (%)
K-means	53.75	63.79
pLSA [3]	89.74	84.46
HpLSA [3]	76.92	72.30
Cas-pLSA [2]	89.70	76.20
DDP-HMM [7]	87.18	85.14
EMD-L1 [15]	92.30	86.40
SparseEMD [16]	89.74	90.00
SRC [10]	92.30	64.41
D-KSVD [9]	92.30	62.70
SCC [12]	89.74	89.83
HSCR	**94.87**	**91.53**

As shown the experimental results in Table 2, it confirmed that the considered HSCR algorithm outperforms the state-of-the-art algorithms using QMUL dataset.

(a) Basketball (b) Handball

Fig. 2. CVBASE '06 dataset

The Sports dataset: As shown in Figure 2, CVBASE '06 dataset depicts complex sports environments both basketball and handball. The data was recorded by a wide view camera in 5 minutes with 366×288 pixels (25 FPS) and 10 minutes with 384×288 pixels (25 FPS) for basketball and handball dataset, respectively. This sports video includes complex movement from various players in the stadium without moving rules. Hence, it is more difficult to distinguish behavior patterns. To discover high -level behavior patterns, we use two t***ypes*** of group activity as high-level behavior patterns: team offense and defense. To obtain the low-level visual features, video data were divided into several video clips every 3 sec. The test dataset consists of 50, 100 video clips for the basketball and handball data, respectively. The quantitative results of the clustering accuracy based on primary high-level behavior patterns are represented in Table 3. The experimental results show that considered HSCR algorithm outperforms all the others.

Table 3. Clustering accuracy (CVBASE '06)

Algorithms	Basketball (%)	Handball (%)
K-means	84.0	65.0
pLSA [3]	64.0	67.0
HpLSA [3]	70.0	72.0
SRC [10]	94.0	87.0
D-KSVD [9]	94.0	87.0
SCC [12]	76.0	70.0
HSCR	**96.0**	**88.0**

Conclusion

This paper considers a HSCR algorithm for complex scene analysis. The HSCR algorithm determines hierarchical behavior pattern using low-level and high-level behavior basis. The discovered HSCR can be capture most primary pattern to analyze behaviors from the complex video scene, while HSCR algorithm tries to removing noise using dimensionality reduction and sparsity in each level and preserving discriminative behavior patterns using hierarchy.

In the experiments, the HSCR has been used for discovering recurrent primary high-level behaviors in various complex video datasets and has been compared with state-of-the-art algorithms. The extensive experimental results show that the considered algorithm achieved good performance for complex scene analysis. In the future, we will extend considered HSCR algorithm for irregular behavior detection.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.NRF-2011-0017202 and NRF- 2010-0028680)

References

1. Y.Yang, J.Liu, and M.Shah, "Video scene understanding using multi-scale analysis", ICCV, pp.1669-1676, 2009.
2. J.Li, S.Gong, and T.Xiang. "Learning behavioural context", IJCV, Vol:97(3), pp.276-304, 2012.
3. J.Li, S.Gong, and T.Xiang, "Global behaviour inference using probabilistic latent semantic analysis ", BMVC, pp.193-202, 2008.
4. X.Wang, X.Ma, and W.Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models", IEEE Trans. on PAMI, Vol:31(3), pp.539-555, 2009.
5. R.Emonet, J.Varadarajan, and J.Odobez, "Extracting and locating temporal motifs in video scenes using a hierarchical non parametric bayesian model", CVPR, pp.3233–3240, 2011.
6. T.Hospedales, S.Gong, and T.Xiang. "A markov clustering topic model for mining behaviour in video", ICCV, pp.1165–1172, 2009.
7. D.Kuettel, M.Breitenstein, L.V.Gool, and V.Ferrari, "What's going on? discovering spatio-temporal dependencies in dynamic scenes", CVPR, pp.1951-1958. 2010.
8. M.Aharon, M.Elad, and A.Brucksteinm, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation", IEEE Trans. on Signal processing, Vol:54(11), pp.4311-4322, 2006.
9. Q.Zhang, B.Li, "Discriminative k-svd for dictionary learning in face recognition", CVPR, pp.2691-2698, 2010.
10. J.Wright, AY.Yang, A.Ganesh, S.Sastry, and Y.Ma,"Robust face recognition via sparse representation", IEEE Trans. on PAMI, Vol:31(2), pp.210-227, 2009.
11. D.Cai, X.He, and J.Han, "Spectral regression for efficient regularized subspace learning", ICCV, pp.1-8, 2007.
12. D.Cai, H.Bao, and X.He,"Sparse concept coding for visual analysis", CVPR, pp.2905-2910, 2011.
13. B.Efron, T.Hastie, I.Johnstone, and R.Tibshirani, "Least Angle Regression", The Annals of Statistics, Vol:32(2), pp.407–499, 2004.
14. G.Zen and E.Ricci, "Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes", CVPR, pp.3225-3232, 2011.
15. G.Zen, E.Ricci, and N.Sebe, "Exploiting sparse representations for robust analysis of noisy complex video scenes", ECCV, pp.193-213, 2012.
16. http://vision.fe.uni-lj.si/cvbase06/index.html.