Journal of Electronic Imaging

JElectronicImaging.org

# Automatic temporal segment detection via bilateral long short-term memory recurrent neural networks

Bo Sun
Siming Cao
Jun He
Lejun Yu
Liandong Li

Bo Sun, Siming Cao, Jun He, Lejun Yu, Liandong Li, "Automatic temporal segment detection via bilateral long short-term memory recurrent neural networks," *J. Electron. Imaging* **26**(2), 020501 (2017), doi: 10.1117/1.JEI.26.2.020501.

# Automatic temporal segment detection via bilateral long short-term memory recurrent neural networks

**Bo Sun, Siming Cao, Jun He,* Lejun Yu, and Liandong Li**
Beijing Normal University, College of Information Science and Technology, Beijing, China

**Abstract.** Constrained by the physiology, the temporal factors associated with human behavior, irrespective of facial movement or body gesture, are described by four phases: neutral, onset, apex, and offset. Although they may benefit related recognition tasks, it is not easy to accurately detect such temporal segments. An automatic temporal segment detection framework using bilateral long short-term memory recurrent neural networks (BLSTM-RNN) to learn high-level temporal–spatial features, which synthesizes the local and global temporal–spatial information more efficiently, is presented. The framework is evaluated in detail over the face and body database (FABO). The comparison shows that the proposed framework outperforms state-of-the-art methods for solving the problem of temporal segment detection.  [DOI: 10.1117/1.JEI.26.2.020501]

Keywords: bilateral long short-term memory recurrent neural networks; temporal segment; temporal–spatial features.

Paper 160985L received Dec. 1, 2016; accepted for publication Feb. 14, 2017; published online Mar. 2, 2017.

## 1 Introduction

Human behavior recognition is an important subject in the field of pattern recognition, which is of great interest for human–computer interaction. Although they are constrained by one's physiology, the temporal factors of human behavior, e.g., facial movements or body gestures, are described by four phases: neutral, onset, apex, and offset.[1] Different phases have different manifestations and intensities. Some researchers have shown that active unit (AU) activation detection[2] and genuine (spontaneous) emotion recognition[3] both benefit from different temporal segments. In addition, they have been proven to be beneficial for emotion recognition, especially for multimodal emotion recognition combining facial expression, body expression, voice, etc.[4–11] Therefore, the temporal segment detection of human behavior warrants further exploration.

The work introduced in this letter offers the following contribution: to date, we are the first to introduce the bilateral long short-term memory recurrent neural networks (BLSTM-RNN) for the automatic detection of the temporal phases of human behavior. A high-level feature that simultaneously contains temporal–spatial information is learned with the BLSTM-RNN method, which has synthesized both local and global temporal information. It shows outstanding performance.

The remainder of this letter is organized as follows. Section 2 introduces related works. Section 3 provides details of the overall methodology. Section 4 describes the experiments and the extensive experimental results. Finally, the conclusion is given in Sec. 5.

## 2 Related Work

In this section, we will review some existing methods that are related to temporal segments. A number of studies have detected the temporal segments by temporal rules drawn up by researchers or other classification schemes, such as support vector machines (SVM)[12] and hidden Markov models (HMMs).[13] Pantic and Patras[14,15] used the temporal rules they drew up to detect the temporal segment of facial AUs. Focusing on bimodal affect recognition, Gunes and Piccardi[10] proposed a method to automatically detect temporal segment of facial movements and body gestures, which includes both frame- and sequence-based strategies. HMM was applied as a sequence-based classifier. Several different algorithms provided in the Weka tool, including SVM, AdaBoost, and C4.5, have been utilized as frame-based classifiers. Jiang et al.[2] used HMMs to detect the temporal segment of facial AUs. Chen et al.[16] design two features to describe face and gesture information, then use SVM to segment expression phase. However, the segment accuracy (Acc) of the existing methods is still to be improved.

Recently, deep learning methods have become very popular within the community of computer vision. BLSTM-RNN is one of the state-of-the-art machine-learning techniques.

In this letter, to synthesize the local and global temporal–spatial information more efficiently, we present an automatic temporal segment-detection framework that uses BLSTM-RNN[17] to learn high-level temporal–spatial features. The framework is evaluated in detail over the face and body (FABO) database.[18] The result of the experiments in Sec. 4 proves that the proposed framework outperforms other state-of-the-art methods for solving the problem of temporal segment detection.

## 3 Methodology

This section presents the details of our method. Figure 1 shows an overview of our proposed method, which consists of two major parts: (1) data preprocessing and (2) feature extraction and representation. We use SVM as a classifier.

### 3.1 *Data Preprocessing*

Face detection is a very important step in the entire pipeline of facial movement temporal segment detection, which directly affects the effectiveness of the feature extraction. Before considering more accurate feature extraction by using methods, such as discrete cosine transform (DCT) combined,[19] the local binary pattern (LBP),[20] the pyramid histogram of oriented gradient (PHOG),[21] the entropy ($E$),[22] the motion area (MA),[16] and the neutral divergence (ND),[16] we follow the methods of[23–26] All frames are aligned to this base face through affine transformation and cut to $200 \times 200$ pixels.

*Address all correspondence to: Jun He, E-mail: hejun@bnu.edu.cn

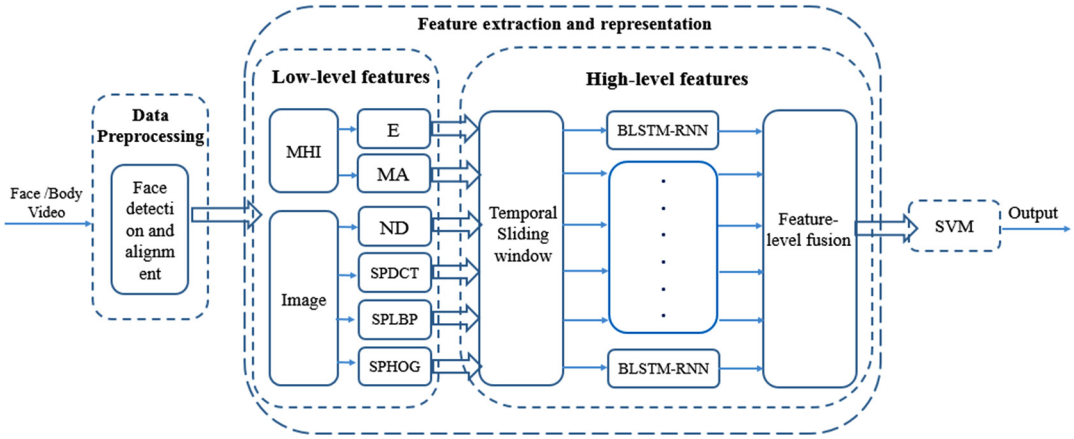


**Fig. 1** Flow chart of the proposed method to detect temporal segments of body gesture and facial movement over the FABO database.

### 3.2 *Feature Extraction and Representation*

#### 3.2.1 *Low-level features*

We extract $E$ and MA based on the motion history image (MHI). An MHI is a static image template which is helpful in understanding the motion location and path as it progresses.[27]

As is common for facial movements, we extract six low-level descriptors, including the sum of pyramid LBP (SPLBP), the sum of pyramid two-dimensional DCT combined (SPDCT), the sum of PHOG (SPHOG), $E$, MA, and ND. For body gestures, only $E$, MA, and ND are extracted as low-level descriptors. Brief introductions can be found in Refs. 16, 19–22.

#### 3.2.2 *High-level features*

After the first step, we obtain several low-level features. Additional temporal information is obtained by learning the high-level features in the time domain with the BLSTM-RNN method and feature-level fusion strategy, which was used to synthesize the local and global temporal information.

Because the input of BLSTM-RNN is required to be a sequence, first, we employ a fixed-size temporal window with its center located at the current frame. Figure 2 shows some examples. In our experiments, we set the temporal window size equal to that of Chen et al.[16] Therefore, each low-level feature vector has the dimensionality of the temporal window size. Then these feature vectors can be used as input for BLSTM-RNN.

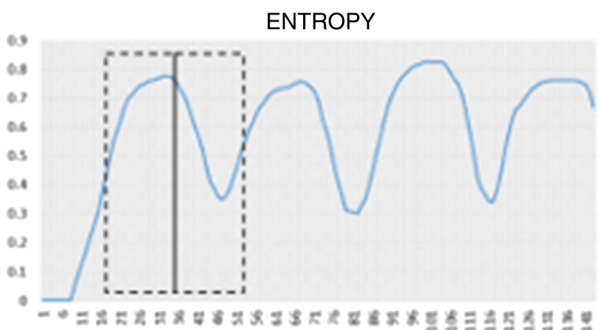


**Fig. 2** Examples of feature representation of current frame, which frames are within a fixed-size temporal window centered at current frame.

For BLSTM-RNN, we use the implementation of Theano.[28] Figure 3 shows the network structure of high-level features. First, we use low-level features as input for BLSTM-RNN and compute the forward hidden sequence $\vec{h}$ and the backward hidden sequence $\overleftarrow{h}$, respectively. Then we link them together by concatenation.

Then we use a feature-level fusion strategy for all of the extracted high-level features. This fusion can be implemented by concatenating all feature vectors together.

Finally, we employ SVM[29] as the classifier to detect temporal phases.

## 4 Experiments

### 4.1 *Experimental Setup*

We conduct the experiments on the bimodal face and body database FABO.[18] This database consists of both face and body recordings using two cameras simultaneously. So far, it is the only bimodal database that has both expression annotation and temporal annotation. We choose $245 \times 2$ videos in which the ground truth expressions from both the face camera and body camera are identical. Among them, 129 videos were used for training and 119 videos were used for testing. In this section, we select the ACC as the measure to evaluate the results. The calculating equation of ACC is given as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s} (TP_i + FP_i)}, \tag{9}$$

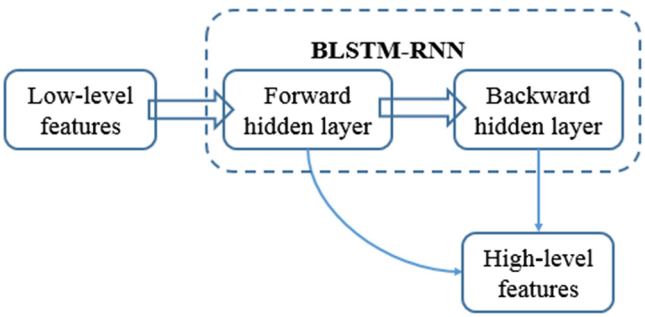


**Fig. 3** Network structure for learning high-level descriptors.

where $s$ denotes the number of temporal phase categories, $P_i$ denotes the precision of the $i$'th temporal phase class, and $TP_i$ and $FP_i$ denote the number of correct classification and the number of wrong classifications in the $i$'th temporal phase class, respectively.

### 4.2 *Experimental Results*

In this section, we compare the ACC in percent under different conditions.

#### 4.2.1 *Result of feature-level fusion*

In this part, we apply three methods for feature-level fusion, and the best one is used on the experimental results.

Table 1 shows the result of feature-level fusion from features of body gestures using BLSTM-RNN with a softmax classifier. In the table, the item BLSTM(1) means using the output of the first layer of BLSTM as features; the item BLSTM(2) means using the output of the second layer of BLSTM as features; the item BLSTM(1+2) means concatenating the output of the first and second layers of BLSTM as final features; the item ALL(before) means first simply concatenating all low-level features together, then combining them for input into a BLSTM-RNN; the item ALL(after) means first use each low-level feature as input into a BLSTM-RNN, then simply concatenate all high-level features together. The results in this table indicate that the most accurate result is 95.20. It is obtained by first using each single feature as input into a BLSTM-RNN, of which the output of the first and second layers is concatenated together as high-level features. Then apply softmax as the classifier.

Table 2 presents the results of feature-level fusion from features of facial movement by using BLSTM-RNN with softmax. This table indicates that the most accurate result is 88.54, which is obtained by first using each low-level feature as input into a BLSTM-RNN, and concatenating the output of the first and second layers, before simply concatenating all high-level features together.

**Table 1** Result of feature-level fusion from features of body gestures using BLSTM-RNN with softmax classifier.

| | ND | MA | E | ALL(before) | ALL(after) |
|---|---|---|---|---|---|
| BLSTM(1) | 90.75 | 80.15 | 76.97 | 93.54 | 95.09 |
| BLSTM(2) | 90.75 | 80.29 | 78.87 | 93.92 | 95.17 |
| BLSTM(1+2) | 90.75 | 80.83 | 79.74 | 94.10 | 95.20 |

**Table 3** Classification results of some classifiers for body gesture and facial movement.

| | ALL (low level) | ALL (high level) |
|---|---|---|
| (a) | | |
| Softmax | 90.86 | 95.20 |
| SVM | 76.30 | 95.30 |
| RF | 85.30 | 94.80 |
| (b) | | |
| Softmax | 87.27 | 88.54 |
| SVM | 72.67 | 89.52 |
| RF | 76.02 | 89.23 |

#### 4.2.2 *Results of some classifiers on low-level features and high-level features*

In this part, we compare different classifiers for temporal segmentation.

Table 3(a) contains the classification results of some classifiers [e.g., softmax, SVM, random forest (RF)] for body gestures. In this table, the item ALL(low level) means simply concatenating all low-level feature vectors together. The item ALL(high level) means simply concatenating all high-level features together. The results in this table show that: (1) the best result is 95.30, which is obtained with the SVM, and (2) the results on high-level features are more accurate than those on low-level features. It shows the validity of high-level features.

Table 3(b) shows the classification results of some classifiers for facial movement. The results in this table show the following: (1) the best result is 89.52, which is obtained with the SVM and (2) the results on high-level features are more accurate than on low-level features. It shows the validity of high-level features. From Table 3, we can see that SVM performs best, so we apply SVM as our final classifier.

#### 4.2.3 *Performance comparison*

In this part, we present a comparison of results from feature- and decision-level fusion for facial movement and body gesture, respectively, in Table 4, and a comparison of results from our approach and relevant experiments on FABO database in Table 5.

Table 4 shows the result of feature- and decision-level fusion for facial movement and body gesture. These results

**Table 2** Result of feature-level fusion from features of facial movement using BLSTM with softmax.

| | SPDCT | *E* | SPLBP | MA | ND | SPHOG | ALL(before) | ALL(after) |
|---|---|---|---|---|---|---|---|---|
| BLSTM(1) | 82.23 | 65.64 | 83.92 | 68.51 | 77.48 | 83.40 | 86.93 | 88.03 |
| BLSTM(2) | 82.39 | 68.56 | 83.74 | 70.39 | 78.27 | 84.27 | 87.73 | 88.12 |
| BLSTM(1+2) | 82.55 | 68.73 | 83.96 | 70.91 | 78.78 | 84.43 | 87.93 | 88.54 |

**Table 4** Result of feature- and decision-level fusion for facial movement and body gesture.

| | Face | Body |
|---|---|---|
| Feature-level fusion | 89.52 | 95.20 |
| Decision-level fusion | 84.56 | 90.62 |

**Table 5** Performance comparison between the proposed and state-of-the-art approaches.[10,16]

| | Face | Body |
|---|---|---|
| Proposed approach | 89.52 | 95.20 |
| Gunes and Piccardi[10] | 57.27 | 80.66 |
| Chen et al.[16] | 83.10 | |

indicate that the results obtained for feature-level fusion are more accurate than those obtained for decision-level fusion. Thus, we employed feature-level fusion as our final strategy.

The performance of the proposed approach and the state-of-the-art approach reported in Refs. 10 and 16 is compared in Table 5. Gunes and Piccardi[10] proposed a method to automatically detect temporal segment of facial movement and body gesture, which includes both frame- and sequence-based strategies. HMM was applied as a sequence-based classifier. Several different algorithms provided in the Weka tool, including SVM, AdaBoost, and C4.5, have been utilized as frame-based classifiers. They obtained an ACC of 57.27% for face and 80.66% for body, respectively. Chen et al.[16] designed two features to describe face and gesture information, then used SVM to segment expression phase. They only used body video to detect the temporal segment of the expression; they did not detect the temporal segment of facial movement and body gesture separately. They obtained an ACC of 83.10% for expression. Results of the proposed approach are obviously more accurate than those of the state-of-the-art methods.

## 5 Conclusions

This letter presents a temporal segment detection framework using BLSTM-RNN to learn high-level temporal–spatial features. The framework is evaluated in detail using data obtained from the FABO database. A comparison with other state-of-the-art methods shows that our method outperforms the other approaches in terms of temporal segment detection. In the future, we plan to focus on affect recognition based on temporal selection face and body display.

### *Acknowledgments*

This work was supported by the National Natural Science Foundation of China (Grant No. 61501035), the Fundamental Research Funds for the Central Universities of China (Grant No. 2014KJJCA15), and the National Education Science Twelfth Five-Year Plan Key Issues of the Ministry of Education (Grant No. DCA140229).

### *References*

1. P. Ekman, "About brows: emotional and conversational signals," in *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, M. V. Cranach et al., Eds., pp. 169–248, Cambridge University Press, New York (1979).
2. B. Jiang et al., "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. Cybern.* **44**(2), 161–174 (2014).
3. H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Recognition of genuine smiles," *IEEE Trans. Multimedia* **17**(3), 279–294 (2015).
4. N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis," *Psychol. Bull.* **111**(2), 256–274 (1992).
5. T. Balomenos et al., "Emotion analysis in man–machine interaction systems," in *Proc. Workshop Multimodal Interaction Related Machine Learning Algorithms*, pp. 318–328 (2004).
6. A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proc. ACM Int. Conf. Multimedia*, pp. 677–682 (2005).
7. C. L. Lisetti and F. Nasoz, "MAUI: a multimodal affective user interface," in *Proc. ACM Int. Conf. Multimedia*, pp. 161–170 (2002).
8. M. Kächele et al., "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," *Int. Conf. on Pattern Recognit. Appl. and Methods* **1**(1), 671–678 , SCITEPRESS - Science and Technology Publications, Lda (2014).
9. M. Valstar et al., "AVEC 2016—Depression, mood, and emotion recognition workshop and challenge," in *Proc. AVEC Workshop* (2016).
10. H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst. Man Cybern. B* **39**(1), 64–84 (2009).
11. S. Piana et al., "Adaptive body gesture representation for automatic emotion recognition," *ACM Trans. Interact. Intell. Syst.* **6**(1), 6 (2016).
12. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.* **20**(3), 273–297 (1995).
13. L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.* **3**(1), 4–16 (1986).
14. M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Trans. Syst. Man Cybern. B* **36**(2), 433–449 (2006).
15. M. Pantic and I. Patras, "Detecting facial actions and their temporal segments in nearly frontal-view face image sequences," in *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, Vol. 4, pp. 3358–3363 (2005).
16. S. Chen et al., "Segment and recognize expression phase by fusion of motion area and neutral divergence features," in *Proc. IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops (FG'11)*, pp. 330–335 (2011).
17. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).
18. H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Proc. Int. Conf. on Pattern Recognition* (2006).
19. N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.* **C-23**(1), 90–93 (1974).
20. T. Ojala and M. Pietikäinen, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002).
21. P. F. Felzenszwalb et al. "Object detection with discriminatively trained part-based models," *IEEE Trans. Software Eng.* **32**(9), 1627–1645 (2010).
22. S. Kirkpatrick, C. D. Gelat, Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science* **220**(4598), 671–680 (1983).
23. K. Sikka, "Multiple kernel learning for emotion recognition in the wild," in *Proc. of the 15th ACM on Int. Conf. on Multimodal Interaction*, pp. 517–524 (2013).
24. A. Dhall et al., "Emotion recognition in the wild challenge 2014: baseline, data and protocol," in *Proc. of the 16th Int. Conf. on Multimodal Interaction*, pp. 461–466 (2014).
25. X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2012)*, pp. 2879–2886 (2012).
26. X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 532–539 (2013).
27. J. W. Davis, "Hierarchical motion history images for recognizing human motion," in *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, pp. 39–46 (2001).
28. F. Bastien et al., "Theano: new features and speed improvements," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning* (2012).
29. E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 130–136 (2000).