Motion and Activity Analysis with Spatiotemporal Local Binary Patterns

Matti Pietikäinen and Guoying Zhao
(mkp,gyzhao)@ee.oulu.fi

Machine Vision Group
University of Oulu, Finland
http://www.ee.oulu.fi/mvg/

Contents

1. Introduction to LBP operators in spatial domain
2. Motion analysis with spatiotemporal LBPs
3. Summary

Local Binary Pattern and Contrast operators


An example of computing LBP and C in a 3x3 neighborhood:

\[
\begin{align*}
\text{example} & = \begin{bmatrix} 6 & 5 & 2 \\ 7 & 6 & 1 \\ 9 & 8 & 7 \end{bmatrix} & \text{thresholded} & = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} & \text{weights} & = \begin{bmatrix} 1 & 2 & 4 \\ 28 & 8 & 0 \\ 64 & 32 & 16 \end{bmatrix} \\
\text{Pattern} & = 11110001 & \text{LBP} & = 1 + 16 + 32 + 64 + 128 = 241 & \text{C} & = (6+7+8+9+7)/5 - (5+2+1)/3 = 4.7
\end{align*}
\]

Important properties:
- LBP is invariant to any monotonic gray level change
- Computational simplicity
Multiscale LBP


- arbitrary circular neighborhoods
- uniform patterns
- multiple scales
- rotation invariance
- gray scale variance as contrast measure

The value of the LBP code of a pixel \((x_c, y_c)\) is given by:

\[
LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p
\]

\[
s(x) = \begin{cases} 
1, & \text{if } x \geq 0; \\
0, & \text{otherwise}.
\end{cases}
\]

1. Sample
2. Difference
3. Threshold
4. Multiply by powers of two and sum

\[1 \times 1 + 1 \times 2 + 1 \times 4 + 1 \times 8 + 0 \times 16 + 0 \times 32 + 0 \times 64 + 0 \times 128 = 15\]
• Bit patterns with 0 or 2 transitions
  \(0 \rightarrow 1\) or \(1 \rightarrow 0\)
  when the pattern is considered circular

• All non-uniform patterns assigned to a single bin

• 58 uniform patterns in case of 8 sampling points
Texture primitives ("micro-textons") detected by the uniform patterns of LBP

Estimation of empirical feature distributions

Input image (region) is scanned with the chosen operator(s), pixel by pixel, and operator outputs are accumulated into a discrete histogram.
Multiscale analysis

Information provided by \( N \) operators can be combined simply by summing up operatorwise similarity scores into an aggregate similarity score:

\[
L_N = \sum_{n=1}^{N} L_n \quad \text{e.g.} \quad LBP_{8,1} \quad + \quad LBP_{8,3} \quad + \quad LBP_{8,5}
\]

Effectively, the above assumes that distributions of individual operators are independent.

Nonparametric classification principle

Sample \( S \) is assigned to the class of model \( M \) that maximizes

\[
L(S, M) = \sum_{b=0}^{B-1} S_b \ln M_b
\]

Many other dissimilarity measures can be used (chi square, histogram intersection, Kullback-Leibler divergence, Jeffrey’s divergence, etc.)

Nonparametric: no assumptions about underlying feature distributions are made!!
Face analysis using local binary patterns

- Face recognition is one of the major challenges in computer vision.
- We proposed (ECCV 2004, PAMI 2006) a face descriptor based on LBP’s.
- Our method has already been adopted by many leading scientists and groups.
- Computationally very simple, excellent results in face recognition and authentication, face detection, facial expression recognition, gender classification.

Face description with LBP


A facial description for face recognition:
Dynamic texture recognition


Dynamic texture

- Dynamic Textures (DT): Temporal texture
- Textures with motion
- An extension of texture to the temporal domain
- Encompass the class of video sequences that exhibit some stationary properties in time

- Lots of dynamic textures in real world
- Description and recognition of DT is needed
Volume Local Binary Patterns (VLBP)

Sampling in volume

Grey-level values

<table>
<thead>
<tr>
<th>142</th>
<th>139</th>
<th>132</th>
<th>135</th>
<th>130</th>
<th>134</th>
</tr>
</thead>
<tbody>
<tr>
<td>118</td>
<td>120</td>
<td>119</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thresholding

Thresholded values

| 1 | 1 | 1 | 3 | 3 | 1 | 1 | 1 |

Multiply Pattern

Weights

<table>
<thead>
<tr>
<th>4</th>
<th>16</th>
<th>64</th>
<th>256</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(001110010111111111) = 11955

LBP from Three Orthogonal Planes (LBP-TOP)

Length of Feature Vector

- Concatenated LBP
- VLBP

P: Number of Neighboring Points

MACHINE VISION GROUP

UNIVERSITY of OULU
DynTex database

- Our methods outperformed the state-of-the-art in experiments with DynTex and MIT dynamic texture databases
Results of LBP from three planes

<table>
<thead>
<tr>
<th>LBP</th>
<th>XY</th>
<th>XZ</th>
<th>YZ</th>
<th>Con</th>
<th>weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,8,8,1,1,1 riu2</td>
<td>88.57</td>
<td>84.57</td>
<td>86.29</td>
<td>93.14</td>
<td>93.43[2,1,1]</td>
</tr>
<tr>
<td>8,8,8,1,1,1 u2</td>
<td>92.86</td>
<td>88.86</td>
<td>89.43</td>
<td>94.57</td>
<td>96.29[4,1,1]</td>
</tr>
<tr>
<td>8,8,8,1,1,1 Basic</td>
<td>95.14</td>
<td>90.86</td>
<td>90</td>
<td>95.43</td>
<td>97.14[5,1,2]</td>
</tr>
<tr>
<td>8,8,8,3,3,3 Basic</td>
<td>90</td>
<td>91.17</td>
<td>94.86</td>
<td>95.71</td>
<td>96.57[1,1,4]</td>
</tr>
<tr>
<td>8,8,8,3,3,1 Basic</td>
<td>89.71</td>
<td>91.14</td>
<td>92.57</td>
<td>94.57</td>
<td>95.71[2,1,8]</td>
</tr>
</tbody>
</table>

Facial expression recognition


- Determine the emotional state of the face
  - Regardless of the identity of the face
Facial Expression Recognition

Mug Shot

Dynamic Information

Action Units

Prototypic Emotional Expressions

[Feng, 2005][Shan, 2005]
[Bartlett, 2003][Littlewort, 2004]
[Tian, 2001][Lien, 1998]
[Bartlett, 1999][Donato, 1999]
[Cohn, 1999][Cohen, 2003]
[Yeasin, 2004][Aleksic, 2005]

Psychological studies [Bassili 1979], have demonstrated that humans do a better job in recognizing expressions from dynamic images as opposed to the mug shot.

(a) Non-overlapping blocks (9 x 8)
(b) Overlapping blocks (4 x 3, overlap size = 10)

(a) Block volumes from three orthogonal planes
(b) LBP features
(c) Concatenated features for one block volume with the appearance and motion
Database

Cohn-Kanade database:

- 97 subjects
- 374 sequences
- Age from 18 to 30 years
- Sixty-five percent were female, 15 percent were African-American, and three percent were Asian or Latino.

Happiness  Angry  Disgust

Sadness  Fear  Surprise
Comparison with different approaches

<table>
<thead>
<tr>
<th></th>
<th>People Num</th>
<th>Sequence Num</th>
<th>Class Num</th>
<th>Dynamic</th>
<th>Measure</th>
<th>Recognition Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Shan, 2005]</td>
<td>96</td>
<td>320</td>
<td>7(6)</td>
<td>N</td>
<td>10 fold</td>
<td>88.4 (92.1)</td>
</tr>
<tr>
<td>[Bartlett, 2003]</td>
<td>90</td>
<td>313</td>
<td>7</td>
<td>N</td>
<td>10 fold</td>
<td>86.9</td>
</tr>
<tr>
<td>[Littlewort, 2004]</td>
<td>90</td>
<td>313</td>
<td>7</td>
<td>N</td>
<td>leave-one-subject-out</td>
<td>93.8</td>
</tr>
<tr>
<td>Tian, 2004</td>
<td>97</td>
<td>375</td>
<td>6</td>
<td>N</td>
<td>-------</td>
<td>93.8</td>
</tr>
<tr>
<td>[Yeasin, 2004]</td>
<td>97</td>
<td>-------</td>
<td>6</td>
<td>Y</td>
<td>five fold</td>
<td>90.9</td>
</tr>
<tr>
<td>[Cohen, 2003]</td>
<td>90</td>
<td>284</td>
<td>6</td>
<td>Y</td>
<td>-------</td>
<td>93.66</td>
</tr>
<tr>
<td>Ours</td>
<td>97</td>
<td>374</td>
<td>6</td>
<td>Y</td>
<td>two fold</td>
<td>95.19</td>
</tr>
<tr>
<td>Ours</td>
<td>97</td>
<td>374</td>
<td>6</td>
<td>Y</td>
<td>10 fold</td>
<td>96.26</td>
</tr>
</tbody>
</table>

Demo for facial expression recognition

- Low resolution
- No eye detection
- Translation, in-plane and out-of-plane rotation, scale
- Illumination change
- Robust with respect to errors in face alignment
Example images in different illuminations

Visible light (VL) : 0.38-0.75 μm
Near Infrared (NIR) : 0.7μm-1.1μm


On-line facial expression recognition from NIR videos

- NIR web camera allows expression recognition in near darkness.
- Image resolution 320 × 240 pixels.
- 15 frames used for recognition.
- Distance between the camera and subject around one meter.

Start sequences  Middle sequences  End sequences
Visual speech recognition


- Visual speech information plays an important role in speech recognition under noisy conditions or for listeners with hearing impairment.

- A human listener can use visual cues, such as lip and tongue movements, to enhance the level of speech understanding.

- The process of using visual modality is often referred to as lipreading which is to make sense of what someone is saying by watching the movement of his lips.

McGurk effect [McGurk and MacDonald 1976] demonstrates that inconsistency between audio and visual information can result in perceptual confusion.
MACHINE VISION GROUP

System overview

Our system consists of three stages.
- First stage: face and eye detectors, and the localization of mouth.
- Second stage: extracts the visual features.
- Last stage: recognize the input utterance.

Local spatiotemporal descriptors for visual information

(a) Volume of utterance sequence
(b) Image in XY plane (147x81)
(c) Image in XT plane (147x38) in y = 40
(d) Image in TY plane (38x81) in x = 70

Overlapping blocks (1 x 3, overlap size = 10).

MOUTH region images
LBP-XY images
LBP-XT images
LBP-YT images
Experiments

- Three databases:

  1) Our own visual speech database: OuluVS Database
     20 persons; each uttering ten everyday’s greetings one to five times.
     Totally, 817 sequences from 20 speakers were used in the experiments.

     | Block volume appearance and motion | LBP features from three ontological planes | Concentrated features for one blocks volume with the |
     |------------------------------------|--------------------------------------------|------------------------------------------------|

     Features in each block volume.

     Mouth movement feature from the whole sequence

     Our own visual speech database: OuluVS Database
     20 persons; each uttering ten everyday’s greetings one to five times.
     Totally, 817 sequences from 20 speakers were used in the experiments.

     | Block volume appearance and motion | LBP features from three ontological planes | Concentrated features for one blocks volume with the |
     |------------------------------------|--------------------------------------------|------------------------------------------------|

     Features in each block volume.

     Mouth movement feature from the whole sequence

     2) Tulips1 audio-visual database
     12 subjects, pronouncing the first four digits in English two times in repetition.
     Totally 96 sequences.

     3) AVLetters database
     10 people, each uttering 26 english letters three times. Totally 780 sequences.
**Experimental results - OuluVS database**

Mouth regions from the dataset.

Speaker-independent:

![Graph showing recognition results for different block volumes.](image)

### Experimental results - Tulips1 audio-visual database

Mouth images with translation, scaling and rotation from Tulips1 database.

Comparison to other methods on Tulips1 audio-visual database (speaker independent).

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Normalization</th>
<th>Results (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Arsic 2006]</td>
<td>MRPCA</td>
<td>Y</td>
<td>81.25</td>
</tr>
<tr>
<td>[Arsic 2006]</td>
<td>MI MRPCA</td>
<td>Y</td>
<td>87.5</td>
</tr>
<tr>
<td>[Gurban 2005]</td>
<td>Temporal Derivatives Features</td>
<td>Y</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>91(a&amp;v, 10 dB SNR level)</td>
</tr>
<tr>
<td>Ours</td>
<td>LBP – TOP_{8,8,1,1,1,3} Blocks: 3x6x2</td>
<td>N</td>
<td>92.71</td>
</tr>
</tbody>
</table>
AVLetters database: 26 letters, 10 people, three utterances per letter.

Principal appearance and motion from boosted spatiotemporal descriptors


Multiresolution features=>Learning for pairs=>Slice selection

- 1) Use of different number of neighboring points when computing the features in XY, XT and YT slices

- 2) Use of different radii which can catch the occurrences in different space and time scales
3) Use of blocks of different sizes to have global and local statistical features

- The first two resolutions focus on the pixel level in feature computation, providing different local spatiotemporal information.
- The third one focuses on the block or volume level, giving more global information in space and time dimensions.

Learned first 15 slices (left) and five blocks (right), each block includes three slices from LBP – TOP8,8,8,3,3,3 with 2 × 5 × 3 blocks for all classes learning.

The selected features for all classes are mainly from YT slices (seven out of 15) and XT slices (seven out of 15), just one from XY slices. That suggests that in visual speech recognition the motion information is more important than the appearance.
Selected 15 slices for phrases “See you” and “Thank you”.

Selected 15 slices for phrases “Excuse me” and “I am sorry”.

These phrases were most difficult to recognize because they are quite similar in the latter part containing the same word “you”. The selected slices are mainly in the first and second part of the phrase.

The phrases “excuse me” and “I am sorry” are different throughout the whole utterance, and the selected features also come from the whole pronunciation.

Demo for visual speech recognition
Face recognition from videos


Problem description

How to efficiently recognize faces, determine gender, estimate age etc. from video sequences?
Traditional approaches..

The most common approach is to apply still image based methods to some selected (or all) frames.

One new direction..

- A Spatiotemporal Approach to Face Analysis from Videos

**Motivations:**

neuropsychological studies indicating that facial dynamics do support face and gender recognition especially in degraded viewing conditions such as poor illumination, low image resolution…
A face sequence can be seen as a collection of rectangular prisms (volumes) from which we extract local histograms of Extended Volume Local Binary Pattern code occurrences.

Algorithm:
1. Divide the video into local prisms
2. Consider 3D neighborhood of each pixel
3. Apply VLBP
4. Feature Selection using AdaBoost
5. Extract local histograms
6. Histogram concatenation & normalization
7. Matching
Some experimental results

Experiments on face recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>Results on MoBo</th>
<th>Results on Honda/UCSD</th>
<th>Results on CRIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>87.1%</td>
<td>69.9%</td>
<td>89.7%</td>
</tr>
<tr>
<td>LDA</td>
<td>90.8%</td>
<td>74.5%</td>
<td>91.5%</td>
</tr>
<tr>
<td>LBP [13]</td>
<td>91.3%</td>
<td>79.6%</td>
<td>93.0%</td>
</tr>
<tr>
<td>HMM [8]</td>
<td>92.3%</td>
<td>84.2%</td>
<td>85.4%</td>
</tr>
<tr>
<td>ARMA [7]</td>
<td>93.4%</td>
<td>84.9%</td>
<td>80.0%</td>
</tr>
<tr>
<td>VLBP [14]</td>
<td>90.3%</td>
<td>78.3%</td>
<td>88.7%</td>
</tr>
<tr>
<td>VLBP+AdaBoost</td>
<td>96.5%</td>
<td>89.1%</td>
<td>94.4%</td>
</tr>
<tr>
<td>EVLBP+AdaBoost</td>
<td>97.9%</td>
<td>96.0%</td>
<td>98.5%</td>
</tr>
</tbody>
</table>
Experiments on gender classification


Databases: CRIM, VidTIMIT and Cohn-Kanade

Gender classification results on test videos of familiar (columns 1-3) and unfamiliar subjects (columns 4-6). The methods are based on appearance only (1st, 2nd & 3rd rows), motion only (4th & 5th rows), and combination of appearance and motion (6th & 7th rows).

<table>
<thead>
<tr>
<th>Method</th>
<th>Gender Classification Rate</th>
<th>Subjects Seen during Training</th>
<th>Subjects Unseen during Training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20 / 20</td>
<td>40 / 40</td>
<td>60 / 60</td>
</tr>
<tr>
<td>Procrustes + SVM + Voting</td>
<td>93.1</td>
<td>93.3</td>
<td>91.9</td>
</tr>
<tr>
<td>LBP + SVM + Voting</td>
<td>94.0</td>
<td>94.4</td>
<td>95.4</td>
</tr>
<tr>
<td>XY-LBP + SVM</td>
<td>96.1</td>
<td>97.2</td>
<td>97.1</td>
</tr>
<tr>
<td>YFLBP + SVM</td>
<td>74.5</td>
<td>81.6</td>
<td>83.2</td>
</tr>
<tr>
<td>YFLBP + SVM</td>
<td>78.5</td>
<td>79.4</td>
<td>80.4</td>
</tr>
<tr>
<td>VLLBP + SVM</td>
<td>96.2</td>
<td>98.3</td>
<td>98.8</td>
</tr>
<tr>
<td>ENLBP + AdaBoost</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Activity recognition

Texture based description of movements

- We want to represent human movement with its local properties
  - Texture
- But texture in an image can be anything? (clothing, scene background)
  - Need preprocessing for movement representation
  - We use temporal templates to capture the dynamics
- We propose to extract texture features from temporal templates to obtain a short term motion description of human movement.

Features

Hidden Markov Models (HMM)

- Model is defined with:
  - Set of observation histograms $H$
  - Transition matrix $A$
  - State priors
- Observation probability is taken as intersection of the observation and model histograms:

$$P(h_{obs} | s_i = q_i) = \sum \min( h_{obs}, h_i)$$
Experiments

- Experiments on two databases:
  - Database 1:
    - 15 activities performed by 5 persons
  - Database 2 - Weizmann database:
    - 10 Activities performed by 9 persons
    - Walking, running, jumping, skipping etc.

Experiments – HMM classification

- Database 1 – 15 activities by 5 people
- LBP
  - \(MHI\) 99%
  - \(MEI\) 90%
  - \(MHI + MEI\) 100%

- Weizmann database – 10 activities by 9 people
- LBP\(_{4,1}\)

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Act</th>
<th>Seq</th>
<th>Res</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>10</td>
<td>90</td>
<td>97.8%</td>
</tr>
<tr>
<td>Wang and Suter 2007</td>
<td>10</td>
<td>90</td>
<td>97.8%</td>
</tr>
<tr>
<td>Boiman and Irani 2006</td>
<td>9</td>
<td>81</td>
<td>97.5%</td>
</tr>
<tr>
<td>Niebles et al 2007</td>
<td>9</td>
<td>83</td>
<td>72.8%</td>
</tr>
<tr>
<td>Ali et al. 2007</td>
<td>9</td>
<td>81</td>
<td>92.6%</td>
</tr>
<tr>
<td>Scovanner et al. 2007</td>
<td>10</td>
<td>92</td>
<td>82.6%</td>
</tr>
</tbody>
</table>
Experiments – Continuous data

- Detection and recognition experiments on database 1 using a sliding window based detection.
  
  Demo

Activity recognition using dynamic textures

- Instead of using a method like MHI to incorporate time into the description, the dynamic texture features capture the dynamics straight from image data.

- When image data is used, accurate segmentation of the silhouette is not needed
  - Instead a bounding box of a person is sufficient!!

Dynamic textures for action recognition

- Illustration of xyt-volume of a person walking

- Formation of the feature histogram for an xyt volume of short duration

  - HMM is used for sequential modeling
Action classification results – Weizmann dataset

- Classification accuracy 95.6% using image data

<table>
<thead>
<tr>
<th>Action</th>
<th>Bend</th>
<th>Jack</th>
<th>Jump</th>
<th>Pjump</th>
<th>Run</th>
<th>Side</th>
<th>Skip</th>
<th>Walk</th>
<th>Wave1</th>
<th>Wave2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bend</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Jack</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Jump</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Pjump</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Run</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Side</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Skip</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Walk</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Action classification results - KTH

- Classification accuracy 93.8% using image data

<table>
<thead>
<tr>
<th>Action</th>
<th>Box</th>
<th>Clap</th>
<th>Wave</th>
<th>Jog</th>
<th>Run</th>
<th>Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box</td>
<td>0.97</td>
<td>0.003</td>
<td>0.987</td>
<td>0.991</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>Clap</td>
<td>0.97</td>
<td>0.004</td>
<td>0.987</td>
<td>0.991</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>Wave</td>
<td>0.97</td>
<td>0.004</td>
<td>0.987</td>
<td>0.991</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>Jog</td>
<td>0.95</td>
<td>0.000</td>
<td>0.987</td>
<td>0.991</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>Run</td>
<td>0.80</td>
<td>0.000</td>
<td>0.987</td>
<td>0.991</td>
<td>0.993</td>
<td>0.999</td>
</tr>
<tr>
<td>Walk</td>
<td>0.00</td>
<td>0.000</td>
<td>0.987</td>
<td>0.991</td>
<td>0.993</td>
<td>0.999</td>
</tr>
</tbody>
</table>
Dynamic textures for gait recognition

$\text{Similarity} = \sum \min(h_i, h_j)$


Experiments - CMU gait database

CMU database
- 25 subjects
- 4 different conditions (ball, slow, fast, incline)
Experiments - Gait recognition results

<table>
<thead>
<tr>
<th></th>
<th>S/B</th>
<th>B/S</th>
<th>F/B</th>
<th>B/F</th>
<th>S/F</th>
<th>F/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMU [4]</td>
<td>92%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>76%</td>
<td>-</td>
</tr>
<tr>
<td>UMD [5]</td>
<td>48%</td>
<td>68%</td>
<td>48%</td>
<td>48%</td>
<td>80%</td>
<td>84%</td>
</tr>
<tr>
<td>MIT [6]</td>
<td>50%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>64%</td>
<td>-</td>
</tr>
<tr>
<td>SSP [7]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>54%</td>
<td>32%</td>
</tr>
<tr>
<td>SVB frieze [8]</td>
<td>77%</td>
<td>89%</td>
<td>61%</td>
<td>73%</td>
<td>82%</td>
<td>80%</td>
</tr>
<tr>
<td>LBP-TOP</td>
<td>75%</td>
<td>83%</td>
<td>75%</td>
<td>83%</td>
<td>88%</td>
<td>88%</td>
</tr>
</tbody>
</table>

Unsupervised dynamic texture segmentation

Dynamic texture segmentation

- Potential applications: Remote monitoring and various type of surveillance in challenging environments:
  - monitoring forest fires to prevent natural disasters
  - traffic monitoring
  - homeland security applications
  - animal behavior for scientific studies.

Related work

- Mixtures of dynamic texture model
  - A.B. Chan and N. Vasconcelos, PAMI2008
- Mixture of linear models
  - L. Cooper, J. Liu and K. Huang, Workshop in ICCV2005
- Multi-phase level sets
  - D. Cremers and S. Soatto, IJCV2004
- Gauss-Markov models and level sets
  - G. Doretto, A. Chiuso, Y. N. Wu and S. Soatto, ICCV2003
- Ising descriptors
  - A. Ghoreyshi and R. Vidal, ECCV2006
- Optical flow
  - R. Vidal and A. Ravichandran, CVPR2005
• Feature: (LBP/C)_{TOP}
  – Local binary patterns
  – Contrast
  – three orthogonal planes

Measure

• Similarity measurement
  \[ \Pi(H_1, H_2) = \sum_{i=1}^{L} \min(H_{1,i}, H_{2,i}) \]

• Distance between two sub-blocks
  \[ d=\{\pi_{LBP,XY}, \pi_{LBP,XT}, \pi_{LBP,YT}, \pi_{C,XY}, \pi_{C,XT}, \pi_{C,YT}\}^T. \]
DT segmentation

- Three phases: Splitting, Merging, Pixelwise classification.

### Splitting

- Recursively split each input frame into square blocks of varying size.
- criterion of splitting:
  - one of the features in the three planes (i.e., LBP_\pi and C_\pi, \pi=XY, XT, YT) votes for splitting of current block.
Merging

- Merge those similar adjacent regions with smallest merger importance (MI) value

  \[ MI = f(p) \times (1 - \Pi) \]
  
  - \( \Pi \) is the distance between two regions
  - \( f(p) = \text{sigmoid}(\beta p) \) (\( \beta = 1, 2, 3, \ldots \))
  - \( p = N_b / N_f \)
  - \( N_b \) is the number of pixels in current block
  - \( N_f \) is the number of pixels in current frame

Pixelwise classification

- Compute \((LBP/C)_{TOP}\) histograms over its circular neighbor for each boundary pixel.

- Compute the similarity between neighbors and connected models.

- Re-label the pixel if the label of the nearest model votes a different label.
Experimental results

Some results on types of sequences and compared with existing methods.

(a) Our method  (b) LBP/C  (c) LBP-TOP  (d) Method in [6]  (e) Method in [7]


Experimental results

• Results on sequences ocean-fire-small

(a) Frame 8  (b) Frame 21  (c) Frame 40

(d) Frame 60  (e) Frame 80  (f) Frame 100
Experimental results


• Results on a real challenging sequence

(a) Frame 5

(b) Frame 10

Dynamic texture synthesis


• Dynamic texture synthesis is to provide a continuous and infinitely varying stream of images by doing operations on dynamic textures.
Introduction

• **Basic approaches to synthesize dynamic textures:**
  - parametric approaches
  - physics-based
  - method and image-based method
  - nonparametric approaches: they copy images chosen from original sequences and depends less on texture properties than parametric approaches

• **Dynamic texture synthesis has extensive applications in:**
  - video games
  - movie stunt
  - virtual reality

Synthesis of dynamic textures using a new representation


- The basic idea is to create transitions from frame $i$ to frame $j$ anytime the successor of $i$ is similar to $j$, that is, whenever $D_{i+1,j}$ is small.
MACHINE VISION GROUP

The algorithm of the dynamic texture synthesis:

1. Frame representation;
   Calculate the concatenated local binary pattern histograms from three orthogonal planes for each frame of the input video

2. Similarity measure;
   Compute the similarity measure $D_{ij}$ between frame pair $I_i$ and $I_j$ by applying Chi-square to the histogram of representation

3. Distance mapping;
   To create transitions from frame $i$ to $j$ when $i$ is similar to $j$, all these distances are mapped to probabilities through an exponential function $P_{ij}$. The next frame to display after $i$ is selected according to the distribution of $P_{ij}$.

4. Preserving dynamics;

5. Avoid dead ends;

6. Synthesis
   Match subsequences by filtering the difference matrix $D_{ij}$ with a diagonal kernel with weights $[w^{-m},...,wm^{-1}]$

Distance measure can be updated by summing future anticipated costs

When transitions of video texture are identified, video frames are played by video loops

Synthesis of dynamic textures using a new representation

An example:

Considering that there are three transitions: $i \rightarrow j_n$ ($n = 1, 2, 3$), loops from the source frame $i$ to the destination frame $j$ would create new image paths, named as loops. A created cycle is shown as:

<table>
<thead>
<tr>
<th>Transitions</th>
<th>$(i_{w}, j_{w})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w=1$</td>
<td>(82, 15)</td>
</tr>
<tr>
<td>$w=2$</td>
<td>(82, 50)</td>
</tr>
<tr>
<td>$w=3$</td>
<td>(67, 23)</td>
</tr>
</tbody>
</table>
Experiments

- We have tested a set of dynamic textures, including natural scenes and human motions. (http://www.texturesynthesis.com/links.htm and DynTex database, which provides dynamic texture samples for learning and synthesizing.)

- The experimental results demonstrate our method is able to describe the DT frames from not only space but also time domain, thus can reduce discontinuities in synthesis. (http://www.ee.oulu.fi/~guoyimo/download/)

Experiments

- Dynamic texture synthesis of natural scenes concerns temporal changes in pixel intensities, while human motion synthesis concerns temporal changes of body parts.

- The synthesized sequence by our method maintains smooth dynamic behaviors. The good performance demonstrates its ability to synthesize complex human motions.
Summary

• Modern texture operators form a generic tool for computer vision
• LBP and its spatiotemporal extensions are very effective for various tasks in computer vision
• Spatiotemporal LBP descriptors combine appearance and motion

• The advantages of the LBP methods include
  - computationally very simple
  - can be easily tailored to different types of problems
  - robust to illumination variations
  - robust to localization errors

• For a bibliography of LBP-related research, see http://www.ee.oulu.fi/research/imag/texture

Example applications using or inspired by spatiotemporal LBPs

- Recognition of dynamic textures: Zhao & Pietikäinen, PAMI 2007
- Segmentation of dynamic textures: Chen et al., ICPR 2008, MLVMA 2009
- Facial expression recognition: Zhao & Pietikäinen, PAMI 2007, PRL 2009; Yang et al., PRL 2009
- Face and gender recognition: Hadid & Pietikäinen, AMFG 2007, PR 2009
- Visual speech recognition: Zhao et al., IEEE T Multimedia 2009
- Background subtraction: Zhong et al., JCIS 2008
- Recognition of actions: Kellokumpu et al., BMVC 2008, MVA 2009
- Recognition of events: Ma & Cisar, ViSU 2009
- Recognition of actions using a sparse descriptor: Mattivi & Shao, CAIP 2009
- Gait recognition: Kellokumpu et al., ICB 2009
- Driver fatigue detection: Yin et al., IJPRAI 2009
- Video texture synthesis: Guo et al., ICIP 2009
Thanks!